Identification and estimation of hidden Markov models

Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultäten

der Philipps-Universität Marburg

vorgelegt von

Grigory Alexandrovich

aus Sankt-Petersburg

Erstgutachter: Prof. Dr. Hajo Holzmann

Zweitgutachter: Prof. Dr. Gunter Ritter

Eingereicht: 19. Mai 2014

Tag der mündlichen Prüfung: 11. September 2014

Contents

Acknowledgments

First of all, I would like to express gratitude to my doctoral supervisor Prof. Dr. Hajo Holzmann for coaching me through PhD and giving a lot of helpful input. I also would like to thank Prof. Dr. Gunter Ritter for making the second assessment of this thesis.

Furthermore I thank my colleagues Viktor Bengs, Dirk Engel, Matthias Eulert, Tobias Filusch, Dr. Daniel Hohmann, Anna Leister, Dr. Florian Ketterer, Dr. Florian Schwaiger and Heiko Werner for the nice time during the last 3.5 years.

Zusammenfassung auf Deutsch

Endliche Mischungsmodelle stellen einen flexiblen Ansatz dar, um heterogene Zufallsphänomene zu modellieren. Heterogene Zufallsphänomene ändern Ihre Eigenschaften je nach Zustand einer latenten Klassenvariablen. Diese latente Klassenvariable kann endlich viele Werte annehmen und ist nicht beobachtbar.

Solche Modelle finden Anwendung in Biologie, Ökonomie, Medizin, Astronomie usw., siehe McLachlan and Peel [32], Titterington et al. [46], Fraley and Raftery [17], McLachlan and Bashford [33].

In der Praxis beliebt sind Mischungen von Gauss-Verteilungen, da sie flexibel und zugleich analytisch zugänglich sind. Für diese Klasse von Mischungen existiert eine Reihe implementierter Verfahren, siehe z.B. Fraley and Raftery [18], R. Lebret [40].

In Kapitel 1 der vorliegenden Dissertation werden einige praktische und theoretische Aspekte der Schätzung einer Gauss-Mischung mittels Maximierung der Likelihood-Funktion betrachtet. Eine Kombination aus dem EM-Algorithmus und dem Newton-Verfahren basierend auf exakten analytischen Ableitungen, wird vorgestellt und mit verschiedenen Implementierungen des EM-Algorithmus in Sektion 1.2 verglichen.

Obwohl der obige Maximum-Likelihood (ML) Ansatz in der Praxis gut funktionert, hat er ein theoretisches Nachteil. Die ML Theorie lässt sich in unserer Situation nicht ohne Weiteres anwenden, da die Likelihood-Funktion einer Gauss-Mischung nicht beschränkt ist.

In der Literatur werden zwei grundlegende Ansätze zur Überwindung des Problems der Unbeschränktheit diskutiert: restringierte Optimierung und Penalisierung der Log-Likelihood. Im ersten Fall wird eine untere Schranke an die Varianzen der Komponenten oder Ihre Quotienten gesetzt, siehe z.B. Hathaway [20]. Im zweiten Fall wird ein Strafterm zur Log-Likelihood addiert, der kleine Varianzen oder Ihre Quotienten penalisiert, siehe z.B. Ciuperca et al. [13], Tanaka [44], Chen et al. [12], Chen and Tan [11]. Der zweite Ansatz hat gegenüber dem Ersten einige Vorteile - es muss keine unbekannte untere Schranke gewählt werden und der Strafterm verschwindet für große Stichproben.

In Sektion 1.3 wird der Beweis für Konsistenz des penalisierten ML-Schätzers für multivariate Gauss-Mischungen aus Chen et al. [12] diskutiert und eine Schwachstelle in der Argumentation identifiziert. Anschliessend wird eine rigorose Korrektur des Beweises auf Grundlage des gleichmäßgien Satzes des iterierten Logarithmus aus Alexander [3] gegeben.

Eine noch flexiblere Klasse von Modellen bilden die sog. hidden Markov models (HMMs), wo eine sequentielle Abhängigkeit ermöglicht und mittels einer latenten Markov-Kette modelliert wird. Anwendungen für diese Modellklasse reichen von Spracherkennung über Sequenzanalyse in Biologie zur Modellierung von Finanzdaten.

Schätzung der Parameter eines HMMs ist ein natürliches statistisches Problem. Leroux [30] zeigte die Konsistenz des ML-Schätzers für parametrische HMMs unter einigen Annahmen. Eine seiner Annahmen ist allerdings verletzt, falls die zustandsabhängigen Verteilungen Gauss-Verteilungen sind, da dann die Likelihood-Funktion, ähnlich wie im Falle von Gauss-Mischungen, unbeschränkt ist. Im Kapitel 2 wird ein zweistufiges Verfahren für konsistente ML-Schätzung von gaussschen HMMs vorgestellt. Im ersten Schritt wird die Dichte der Marginalmischung des HMMs geschätzt indem eine penalisierte Mischungslikelihood maximiert wird. Im zweiten Schritt wird die volle HMM-Likelihood über einer Umgebung der Schätzwerte aus dem ersten Schritt maximiert. Der Konsitenzbeweis ähnelt in seiner Struktur dem Beweis aus Chen et al. [12] für Konsistenz vom penalisierten ML-Schätzer für Gauss-Mischungen. Die Bernstein-Ungleichung aus Merlevède et al. [35] spielt eine wichtige Rolle um eine ähnliche Aussage wie das Lemma 1 aus Chen et al. [12] für HMMs zu erhalten. Diese Aussage ist der entscheidende Punkt im Beweis.

Eine weitaus größere Flexibilität ermöglicht nichtparametrische Modellierung der zustandsabhängigen Verteilungen. Eine wichtige Fragestellung in diesem Zusammenhang ist Identifikation, d.h. die Frage ob die Verteilung des Prozesses $(Y_t)_{t\in\mathbb{N}}$ die zustandsabhängigen Verteilungen und die Übergangsmatrix in einer geeigneten Weise determinieren. Im Kapitel 3 betrachten wir das Problem und beweisen Identifizierbarkeit under den relativ schwachen Annahmen, dass die Übergangsmatrix regulär ist und die zustandsabhängigen Verteilungen unterschiedlich sind. Das Hauptwerkzeug hierbei ist das Resultat von Kruskal [28] zur Identifikation von Faktoren eines drei-dimensionalen Arrays und die Methodik aus Allman et al. [6].

Introduction

Finite mixture models provide a powerful approach for modeling heterogeneous data. Heterogeneity in this case means that the data-generating process consists of several sub-populations.

There is a wide range of applications for mixture models in biology, economics, medicine, astronomy etc., see McLachlan and Peel [32], Titterington et al. [46], Fraley and Raftery [17], McLachlan and Bashford [33].

Cluster analysis is an important domain of application for mixture models. Here the objective is to find clusters, i.e. homogeneous subsets, in the data. In the simplest case of clustering via a mixture model, a mixture density is estimated from the data and each mixture component is associated with a cluster through maximizing the a-posteriori probabilities. Also refinements of this approach, such as merging of weakly separated components, are possible, see e.g. Baudry et al. [7], Hennig [21].

In Chapter 1, practical and theoretical aspects of the estimation of Gaussian mixture models via likelihood maximization will be considered. A combination of the EM algorithm and Newton's method, based on exact analytical derivatives will be introduced and compared with several EM implementations in Section 1.2.

Although the above methods work well in practice, they lack theoretical justification. The maximum likelihood theory is not applicable to Gaussian mixtures without further ado, since the likelihood function is not bounded. Two basic strategies for overcoming the unboundedness were studied in the literature: restricted optimization and penalization of the likelihood. In the first case a lower bound on the variances or their ratios is imposed; see e.g. Hathaway [20]. In the second case a term which penalizes small variances or ratios of variances is added to the log-likelihood; see e.g. Ciuperca et al. [13], Tanaka [44], Chen et al. [12], Chen and Tan [11]. The second approach has some advantages over the first one - there is no tuning constant to choose and the penalty function actually disappears with increasing sample size.

In Section 1.3 we discuss the consistency proof of penalized maximum likelihood estimator for multivariate Gaussian mixture from Chen and Tan [11], identify a

soft spot in the proof and introduce a rigorous correction based on a uniform law of the iterated logarithm from Alexander [3].

An even richer class of models, than finite mixtures are hidden Markov models (HMMs), where a sequential dependence between observations is allowed and modelled by an underlying Markov chain. Applications for these models range from speech recognition over biological sequence analysis to modelling financial data. HMMs with a finite state space and a discrete time will be considered in this thesis.

Estimation of paramertric HMMs is a natural statistical problem. Leroux [30] showed consistency of the maximum likelihood estimator of a parametric HMM under some general assumptions. His assumptions are violated in the case where the state-dependent distributions are Gaussians, so that the MLE does not exist in that case, similar to the situation with Gaussian mixtures. In Chapter 2, a penalized two-step procedure is introduced as a solution of this problem. In the first step the state-dependent distributions are estimated through maximizing a penalized mixture likelihood and in the second step, full HMM likelihood is maximized in a neighborhood of the values from the first stage.

Parametric estimation theory for finite-state HMMs has been well developed within the last two decades. As mentioned above, Leroux [30] obtained consistency of the maximum likelihood estimator, and Bickel et al. [8] and others later the asymptotic normality. Consistency is based on parametric identification of the transition matrix and the parameters of the state-dependent distributions, which Leroux [30] proved from a result by Teicher [45] on identifiability of mixtures of product distributions by considering the joint distribution of two successive observations.

In order to achieve greater flexibility and to avoid model misspecification, nonparametric modeling of the component distributions may be of some interest. However, the first and most basic question is whether such models are still identified, i.e. whether the distribution of the observed layer of a HMM $(Y_t)_{t\in\mathbb{N}}$ determines the state-dependent distributions and the transition probability matrix (t.p.m.) in an appropriate sense. In Chapter 3, we consider this problem and prove identification of nonparameteric HMMs under the assumption that the transition probability matrix is regular and the state-dependent distributions are distinct. The main tool therefor will be identification result for a simple latent-class model from Kruskal [28] and ideas from Allman et al. [6].

All notations in the current thesis are either defined before the first use or are defined in the list of notations at the end.

1 Estimation of finite multivariate Gaussian Mixtures

1.1 Historical overview

The most popular mixture models are finite mixtures of Gaussians. The reasons for this are analytical tractability paired with a high flexibility of the resulting model. A d-dimensional Gaussian mixture model with K components is given by the density

$$g(y;\theta) = \sum_{i=1}^{K} p_i \varphi(y;\mu_i,\Sigma_i),$$

where $\varphi(\cdot;\mu,\Sigma)$ is density of a Gaussian with mean μ and covariance matrix Σ and θ the parameter $(\mu_1,\ldots,\mu_K,\Sigma_1,\ldots,\Sigma_K,p_1,\ldots,p_K) \in \mathbb{R}^d \times \ldots \times \mathbb{R}^d \times \mathcal{P}^d \times \ldots \times \mathcal{P}^d \times \triangle^{K-1}$.

Identification of multivariate Gaussian mixtures up to label swapping was proved by Jakowitz and Spargins [23]. The parameter θ is often estimated through the likelihood maximization (MLE) via Expectation-Maximization algorithm (EM), see Dempster et al. [14], Redner and Walker [42], McLachlan and Krishnan [34], McLachlan and Bashford [33], Xu and Jordan [49]. Despite of theoretical inconsistency of the MLE for Gaussian mixtures due to the unboundedness of the likelihood, which will be treated later, this approach works well in practice.

Although EM has many advantages, it has also several drawbacks, such as mere linear convergence and a bad behavior at a presence of a high fraction of unobserved information. This is why many authors considered improvements of EM or alternative approaches, see Celeux et al. [10], Everitt [15], Aitkin and Aitkin [1], Peters and Walker [38], Jank [25].

Everitt [15] compared six different algorithms for calculation of the MLE of a two-component univariate Gaussian mixture (GM). In particular he compared the EM algorithm, variants of Newton's method with approximated and exact gradient and Hessian, Fletcher-Reeves algorithm and the Nedler-Mead simplex algorithm and

concluded that the most satisfactory algorithms are EM and NM with exact gradient and Hessian. Aitkin and Aitkin [1] also considered a hybrid EM/Newton's algorithm, which starts with five EM iterations and then switches to Newton's method, if Newton's method yields a descent direction, another five EM iterations are done and so on. They used this approach for estimating a MLE of a two-component univariate GM and reported a superior behaviour of the hybrid algorithm over the pure EM.

Peters and Walker [38] developed an iterative procedure for calculating the MLE of a multivariate GM. Their approach is based on the so called likelihood equations, which follow from the critical equation. They constructed a locally contractive operator and obtained a fixed point problem. In the proof of the contractibility they calculate the exact derivatives of the operator. Unfortunately they did not compare their method with other algorithms and it is hard to judge how well it works. No available implementation of their method is known to us.

Lange [29], proposed a quasi-Newton acceleration of the EM algorithm, where given an estimate θ_k, the Hessian of the observed log-likelihood is approximated by a decomposition into the Hessian of the conditional expectation of the complete log-likelihood ($Q(\theta \mid \theta_k)$ from the E-step) minus a part that is constructed via rank-one updates. The gradient of the observed log-likelihood is approximated by the gradient of $Q(\theta \mid \theta_k)$. In accordance with Lange, such an approach yields a faster converging sequence. Jamshidian and Jennrich [24] also considered several acceleration methods of the EM algorithm, in which they used a Quasi-Newton approach among others. They found examples where the accelerated versions are dramatically faster than the pure EM algorithm.

Xu and Jordan [49] discussed the properties of the EM algorithm for calculation of the MLE for GMs. They proved a superiority statement of the EM algorithm over the constrained gradient ascent in a setting with known component weights and covariance matrices and conjectured that it holds also in a general setting. In the numerical experiments they demonstrated a general superiority of EM over the constrained gradient ascent. In their remarks Xu and Jordan speak against the use of NM for estimating GMs due to computational costs and numerical considerations. However they did not compare NM with EM in the numerical experiments and they did not address the problem of high proportion of unobserved information in GM.

The next section is addressed to the calculation of the MLE of θ based on i.i.d. observations $Y_1, \ldots, Y_n$ via a combination of EM algorithm and exact Newton's method. Exact means here the use of analytical derivatives of the log-likelihood with respect to an appropriate parameterization.

1.2 Newton's Method

1.2.1 Parameterization and algorithm

The log-likelihood function of a Gaussian mixture model for an i.i.d. sample $Y_1, \ldots, Y_n$ is given by

$$l(\theta) = \log \prod_{t=1}^{n} g(Y_t; \theta) = \sum_{t=1}^{n} \log g(Y_t; \theta).$$

As indicated before, we apply Newton's method to maximize $l(\theta)$. First of all, we have to find an appropriate parameterization of the mixture g. To ensure that the weights $p_1, \ldots, p_K$ stay in the interval $[0, 1]$ and sum to one during the iterations, we parameterize them as follows:

$$p_i = p_i(q) := \frac{q_i^2}{q_1^2 + \ldots + q_{K-1}^2 + 1}, \quad 1 \leq i \leq K-1, \tag{1.1}$$

where $q = (q_1, \ldots, q_{K-1}) \in \mathbb{R}^{K-1}$. With this approach we avoid optimization under the K inequality restrictions: $p_i \geq 0,\ 1 \leq i \leq K-1,\ \sum_{i=1}^{K-1} p_i \leq 1$.

Regular covariance matrices and their inverses are s.p.d (symmetric and positive definite) matrices, so they can be written as

$$\Sigma_i^{-1} = L_i L_i' \tag{1.2}$$

with $L_i \in \mathbb{R}_{lt}^{d \times d}$ via the Cholesky-decomposition. The only requirement on L_i is that it has only non-zero elements on the diagonal.
From this point on, we parameterize the family of multivariate normals by μ and L:

$$\phi(Y; \mu, L) = \frac{1}{\sqrt{2\pi}^D} |L| e^{-\frac{1}{2}(Y-\mu)' LL' (Y-\mu)},$$

and set $\Sigma^{-1} = LL'$. The mixture density becomes

$$g(x; \mu_1, \ldots, \mu_K, L_1, \ldots, L_K, q_1, \ldots, q_{K-1}) = \sum_{i=1}^{K} p_i(q) \phi(Y; \mu_i, L_i). \tag{1.3}$$

Next, we give a brief introduction to Newton's method for maximizing a twice-differentiable function $f : U \to V$, where $U \subset \mathbb{R}^d, V \subset \mathbb{R}$ for some $d \in \mathbb{N}$. The

essence of the approach is to find an appropriate root of the equation $\nabla_\theta f(\theta) = 0$. In our case the log-likelihood function l will play the role of f. For a more detailed overview of Newton's method see e.g. Kelley [26] or Nocedal and Wright [36].

The Basics. Newton's method is an iterative procedure, that constructs a sequence $(\theta_k)_{k\in\mathbb{N}}$, which converges towards the solution θ^*. The iteration is defined by

$$\theta_{k+1} = \theta_k + t_k \Delta_k, \tag{1.4}$$

where

$$\Delta_k := -H_k^{-1} \nabla_\theta f(\theta_k). \tag{1.5}$$

H_k is the Hessian of f evaluated at θ_k and t_k is a positive step size at the iteration k. We can consider Δ_k as the maximizer of a quadratic approximation of f in θ_k: $f(\theta_k + p) \approx f(\theta_k) + \nabla_\theta f(\theta_k)' p + \frac{1}{2} p' H_k p$. The quality of such an approximation depends on the length of p and the smoothness of f in the neighbourhood of θ_k.

The iteration ends as soon as some convergence criterion is fulfilled, e.g. $||\Delta_k|| \leq \epsilon$ for some small ϵ, or the maximal number of iterations is achieved. To start the iterations one has to supply an appropriate starting point θ_1. The selection of the starting point may be a hard problem, since the neighbourhood of θ^* where Newton's method converges may be very small. One possibility to find a starting point, is to prefix another algorithm such as a gradient method or as in our case the EM algorithm. In a sufficiently small neighbourhood of θ^* Newton's method has a quadratic convergence rate, meaning that $||\theta_{k+1} - \theta^*|| \leq c||\theta_k - \theta^*||^2$ for a $c > 0$.

Line Search. Given a direction Δ_k we need to decide how deep to follow it, i.e. to select an appropriate step length t_k, see (1.4). Since we want to maximize a function, namely the log-likelihood function, one suitable choice would be

$$t_k := \underset{t>0}{\operatorname{argmax}}\, f(\theta_k + t\Delta_k).$$

An exact solution of this problem is often difficult to obtain, so one tries to find an approximation. To achieve a sufficient increase of the objective function, the step length t_k must satisfy the so called Wolfe conditions:

$$\begin{aligned} f(\theta_k + t_k\Delta_k) &\geq f(\theta_k) + c_1 t_k \nabla_\theta f(\theta_k)' \Delta_k \\ \nabla_\theta f(\theta_k + t_k\Delta_k)' \Delta_k &\leq c_2 f(\theta_k)' \Delta_k, \end{aligned}$$

with $0 < c_1 < c_2 < 1$. The constant c_1 is often chosen quite small, near 10^{-4}.

The first inequality is sometimes called Armijo condition and ensures that f will make a sufficient increase along the direction Δ_k and the second condition ensures that the step size t_k will be not too small. In practice one often uses the so called backtracking approach to find an appropriate step length. So do we in our implementation. For a more detailed explanation we again refer to Nocedal and Wright [36].

Solving for Δ_k. At every iteration we have to solve the following system of linear equations for Δ_k:

$$H_k \Delta_k = -\nabla_\theta f(\theta_k).$$

The matrix H_k is symmetric, but not necessarily positive definite. We use the so-called rational Cholesky decomposition $H_k = C_k D_k C_k{}'$ with lower triangular matrix C_k and diagonal matrix D_k. It is called rational, since no roots have to be calculated and it works even if some elements in D_k are negative.

As mentioned above, we apply Newton's method to find an appropriate root of the equation $\nabla_\theta l(\theta) = 0$, where l is the log-likelihood function of the mixture and θ is the parameter vector. For this purpose we need the gradient and the Hessian of the log-likelihood, which are given in the Section 1.2.5.

Penalization. In this paragraph we subscript the log-likelihood and the penalty function to be defined with n, to indicate that they depend on the sample size.

The log-likelihood function of a Gaussian mixture is unbounded. To see this, we observe that for a parameter θ with $\mu_1 = Y_1$, $|\Sigma_1| = \varepsilon$ and μ_2, $\Sigma_2 > 0$ arbitrary, the log-likelihood diverges to ∞ for $\epsilon \to 0$. So it may happen that the algorithm converges toward such a solution. In order to avoid such bad solutions one can penalize the log-likelihood with an additive term $s_n(\theta) = s_n(L_1, \ldots, L_K)$, where n is the sample size. Chen and Tan [11] formulate conditions that must be satisfied by s_n in order to make the resulting estimator consistent. In the next section we will discuss these conditions and the consistency proof of penalized MLE of Chen and Tan and identify a flawed argument therein. A rigorous correction will be given.

A function which satisfies necessary conditions is

$$s_n(L_1, \ldots, L_K) = -a_n \sum_{i=1}^{K} \big(\operatorname{tr}(S L_i L_i{}') + \log \frac{1}{|L_i|^2}\big),$$

where S is the sample covariance matrix and $a_n \to 0$ (e.g. $a_n = \frac{1}{n}$ or $a_n = \frac{1}{\sqrt{n}}$). The penalized log-likelihood has now the form

$$pl_n(\theta) = l_n(\theta) + s_n(\theta). \tag{1.6}$$

If penalization is enabled, the log-likelihood function is replaced by its penalized version. The derivatives of this function arise as the sum of the derivatives of the summands, both are given in Section 1.2.5.

1.2.2 Fisher information

At every iteration of Newton's method we obtain the Hessian of the log-likelihood function $\nabla^2_\theta l(\theta_k)$. A well known result from the MLE framework is that under certain conditions (Redner and Walker, 1984)

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z, \qquad Z \sim N(0, I^{-1}_{\theta_0}),$$

where $\hat{\theta}_n$ is a root of the gradient of the log-likelihood function based on n observations and $I_{\theta_0} = \mathbb{E}\nabla_\theta \log g(Y;\theta_0)\nabla_\theta \log g(Y;\theta_0)' = -\mathbb{E}\nabla^2_\theta \log g(Y;\theta_0)$, the Fisher information matrix. An approximation of $\mathbb{E}\nabla^2_\theta \log g(Y;\theta_0)$ is given by $-\hat{I}_{\hat{\theta}_n} = \frac{1}{n}\sum_{t=1}^{n} \nabla^2_\theta \log g(Y_t;\hat{\theta}_n)$. The last term is the Hessian of the log-likelihood multiplied by $\frac{1}{n}$. The covariance matrix of Z allows us to construct confidence sets for θ_0.

The parameters of interest are $\theta_{int} = (\mu_i, \Sigma_i, p_i;\ i = 1 \dots K)$, we however obtain the Hessian w.r.t $\theta_{new} = (\mu_i, L_i, q_i;\ i = 1 \dots K)$ as defined at the beginning of the section. Let ψ be the map with $\psi(\theta_{new}) = \theta_{int}$ and D_ψ its derivative matrix. By the chain rule the Fisher information matrix for θ_{int} is $D^{-1}_{\psi'} I_{\theta_0} D^{-1}_{\psi}$. ψ is identity in $\mu_1, \dots, \mu_k$. The partial derivatives of ψ w.r.t. q_i and L_i are given in Section 1.2.5.

1.2.3 Numerical comparison to EM

The computation of the gradient and the Hessian of the log-likelihood is the most expensive part of the algorithm and grows linearly with the sample size. This step and the solver for linear equations were implemented in C, since a direct implementation in R was too slow.

An advantage of deriving the log-likelihood is that we have to deal with sums:

$$\nabla_\theta l(\theta) = \sum_{t=1}^{n} \frac{1}{g(Y_t;\theta)} \nabla_\theta g(Y_t;\theta)$$

resp.

$$\nabla^2_\theta l(\theta) = \sum_{t=1}^{n} \frac{1}{g(Y_t;\theta)} \left(\nabla^2_\theta g(Y_t;\theta) - \frac{1}{g(Y_t;\theta)} \nabla_\theta g(Y_t;\theta) \nabla'_\theta g(Y_t;\theta) \right).$$

The summands can be computed in parallel. We used the OpenMP API in our C-Code for this purpose. The parallelized version is available only for Unix OS.

Algorithms. We compared our algorithm with the EM-implementations from the R packages **Mclust** 3.4.11. and **Rmixmod** 1.1.3.. In addition we considered the SEM algorithm, which is also contained in the package Rmixmod. The SEM algorithm is a stochastic version of the EM algorithm, where in each iteration the unobserved variables (the cluster labels z) are drawn from the conditioned density $g_z(z|x;\theta_k) = g_c(x,z;\theta_k)/g(x;\theta_k)$ and then the simulated complete likelihood g_c is maximized. This algorithm was designed to overcome the drawbacks of the EM algorithm, such as convergence towards saddle points and slow convergence rate. See [10] for a more detailed explanation.

The interesting characteristics were the execution time, the accuracy of the solution, measured by the BIC values and the number of the iterations. The BIC (Bayesian information criterion) of θ is given by $2l(\theta) - k \log n$, where k is the number of free parameters. In our case k was fixed, so we essentially compared the achieved values of the log-likelihood.

The initial solution for Newton's method was found by a k-means clustering, followed by the EM algorithm, which terminated as soon as the relative log-likelihood change $\frac{l(\theta_{k+1})-l(\theta_k)}{l(\theta_k)}$ fell below 1e-6. The succeeding Newton's method terminated as soon as one of the following criteria was fulfilled:

C1. The number of iterations achieved 10.

C2. The Hessian of the log-likelihood became singular.

C3. No positive step length was found during back-tracking.

C4. The norm of the Newton's direction Δ_k and the norm of the gradient of the log-likelihood fell below 1e-12.

Newton's method was tested with enabled penalization ($a_n > 0$) and without it ($a_n = 0$). In order to determine the effect of the parallelization, we also tested the parallel version.

The initial solution for the EM algorithm from the package Mclust was the same k-means clustering, which was used for Newton's method, and the algorithm terminated as soon as the relative log-likelihood change fell below 1e-8 (default value in the package Mclust). The method of initialization of the EM/SEM algorithms from the package Rmixmod was an internal random start, since there was no possibility to supply an initial solution. The termination rule was the same as for the EM algorithm from the package Mclust. No restrictions on the parameters were made. We use the following abbreviations for the considered algorithms:

- `NM = Newton's method without penalization`
- `NMP = Newton's method with penalization`
- `EMC = EM algorithm from the package Mclust`
- `EMIX = EM algorithm from the package Rmixmod`
- `SEM  = SEM algorithm from the package Rmixmod`

Procedure. All experiments were realized on a benchmark machine with 12 Intel Xeon X5675 3.07GHz CPUs and 24Gb RAM. We compared the algorithms for five different models, which mimicked some relevant (but of course not all) situations which may occur in the practice. The procedure was the following:

1. Generate N data points from a model.
2. Calculate the MLE with Newton's method (penalized and unpenalized).
3. Calculate the MLE with the EM and SEM algorithms (Mclust and Rmixmod).
4. Save the corresponding numbers of iterations, execution times and BICs.

We repeated this procedure 1000 times and obtained thereby samples of the execution times, iterations numbers and BICs for all algorithms.

Comparing results. The results were evaluated by pairwise comparisons of the samples of the execution times and the BIC values of the algorithms. In order to compare two samples the Wilcoxon-Mann-Whitney test was used.

Given two algorithms A and B, and the corresponding time samples t_A, t_B and the BIC samples BIC_A, BIC_B the following four hypotheses and corresponding p-values were considered:

1. $H^t_{A\lesssim B} : t_A \geq t_B$ and $p^t_{A\lesssim B}$,
2. $H^{BIC}_{A\lesssim B} : BIC_A \leq BIC_B$ and $p^{BIC}_{A\lesssim B}$,
3. $H^t_{A\gtrsim B} : t_A \leq t_B$ and $p^t_{A\gtrsim B}$,
4. $H^{BIC}_{A\gtrsim B} : BIC_A \geq BIC_B$ and $p^{BIC}_{A\gtrsim B}$.

There is a significant advantage of A over B in terms of $crit \in \{t, BIC\}$ if we can reject the hypothesis $H^{crit}_{A\lesssim B}$. We assume that there is no significant advantage of B over A if we cannot reject the hypothesis $H^{crit}_{A\gtrsim B}$. These thoughts lead us to the following definition of a benchmark $\text{bench}(A, B) \in \{-1, 0, 1, 2\}$.

$$\begin{aligned}\text{bench}(A,B) := &- \mathbf{1}_{\{p^t_{A\gtrsim B}\leq 0.05 \,\wedge\, p^{BIC}_{A\gtrsim B}\leq 0.05\}} \\ &+ \mathbf{1}_{\{p^t_{A\lesssim B}\leq 0.05 \,\wedge\, p^{BIC}_{A\gtrsim B}\geq 0.05\}} \\ &+ \mathbf{1}_{\{p^{BIC}_{A\lesssim B}\leq 0.05 \,\wedge\, p^t_{A\gtrsim B}\geq 0.05\}} \\ &+ 2\cdot \mathbf{1}_{\{p^t_{A\lesssim B}\leq 0.05 \,\wedge\, p^{BIC}_{A\lesssim B}\leq 0.05\}}.\end{aligned}$$

In words, we set $\text{bench}(A, B)$ to -1 if A was both significantly slower and significantly worse (in terms of BIC) than B. We set $\text{bench}(A, B)$ to 1 if *either* A was significantly faster than B and at the same time was not significantly worse *or* if A was significantly better than B and at the same time was not significantly slower. Furthermore, if *either* A was significantly faster and significantly worse than B *or* if A was significantly better and significantly slower than B we set $\text{bench}(A, B)$ to 0. Finally, we set $\text{bench}(A, B)$ to 2 if A was significantly faster and significantly better than B.

A higher $\text{bench}(A, B)$ implies an advantage of A over B in a given model/sample size constellation. However $\text{bench}(A, B)$ never achieved 2 in our simulations, since there were no significantly differences in BIC. Some tables with benchmark values are given in Section 1.2.5.

Convergence failures. Especially Newton's method fails to converge if the initial solution is not well chosen. We say the algorithm A failed to converge if either a fatal numerical error occurred (e.g. such as an attempt to invert a singular matrix) or the solution was an extreme outlier. We say a result of the algorithm A is an extreme outlier if and only if the distance between the corresponding BIC value and the median of the BIC sample is greater than 3 times the interquartile range of the sample. Such atypical BIC values correspond to saddle points or local maxima of the log-likelihood with very poor parameter estimates.

All such failure cases were removed from the data before we compared the algorithms. The counts of such failures for each model and each sample size are given in Section 1.2.5 in Table 1.1.

Models. Following models will be considered: **Model 1** $K = 2,\ D = 3$.

$$\begin{aligned}
&\mu_1 = (1\ \ 1\ \ 1)', \quad \mu_2 = (2\ \ 2\ \ 2)' \\
&\Sigma_1^{i,j} = \Sigma_2^{i,j} = \begin{cases} 0.2 & i \neq j \\ 1 & i = j \end{cases} \quad 1 \leq i, j \leq D, \\
&p_1 = p_2 = 0.5.
\end{aligned}$$

This model is interesting, since we are in $\mathbb{R}^3$ and the corresponding covarince matrices are dense, so we have a relatively complex correlation structure within the components and a moderate number of paramters to estimate.

Model 2 $K = 5,\ D = 2$.

$$\begin{aligned}
&\mu_1 = (4\ \ 5)', \quad \mu_2 = (1.5\ \ 5)', \mu_3 = (2\ \ 4.5)', \quad \mu_4 = (4.1\ \ 1)', \quad \mu_5 = (5\ \ 1)', \\
&\Sigma_1 = \begin{pmatrix} 0.3 & 0.05 \\ 0.05 & 0.3 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.2 \end{pmatrix}, \\
&\Sigma_{4,5} = \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}, p_1 = p_4 = p_5 = 0.2, \quad p_2 = 0.25, \quad p_3 = 0.15.
\end{aligned}$$

The interesting characteristics of this model are the high number of components and a strong overlap between components 2 and 3, and 4 and 5, the corresponding 2-component mixtures are unimodal or weakly bimodal respectively.

Model 3 $K = 2,\ D = 5.$

$$\mu_1 = (1\ \ 1\ \ 1\ \ 1\ \ 1)', \quad \mu_2 = (2\ \ 2\ \ 2\ \ 2\ \ 2)',$$
$$\Sigma_1^{i,j} = \Sigma_2^{i,j} = \begin{cases} 0.2 & i \neq j \\ 1 & i = j \end{cases} \ 1 \leq i, j \leq D,$$
$$p_1 = p_2 = 0.5.$$

We include this model in our consideration, since it is quite high dimensional and it is interesting to study the behavior of Newton's method in higher dimensions, but with small number of components.

Model 4 $K = 7,\ D = 2.$

$$\mu_1 = (4\ \ 5)', \quad \mu_2 = (1.5\ \ 5)', \quad \mu_3 = (2\ \ 4.5)', \quad \mu_4 = (4.1\ \ 1)', \mu_5 = (5\ \ 1)',$$
$$\mu_6 = (3\ \ 2)', \quad \mu_7 = (5\ \ 2)',$$
$$\Sigma_1 = \begin{pmatrix} 0.3 & 0.05 \\ 0.05 & 0.3 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.2 \end{pmatrix},$$
$$\Sigma_{4,5,6,7} = \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}$$
$$p_1 = 0.2, \quad p_2 = p_3 = 0.15, \quad p_4 = p_5 = 0.1, \quad p_6 = p_7 = 0.15.$$

Model 5 $K = 9,\ D = 2.$

$$\mu_1 = (4\ \ 5)', \quad \mu_2 = (3\ \ 5)', \quad \mu_3 = (2\ \ 4.5)', \quad \mu_4 = (4.1\ \ 1)', \mu_5 = (5\ \ 1)',$$
$$\mu_6 = (3\ \ 2)', \quad \mu_7 = (5\ \ 2)', \quad \mu_8 = (-1\ \ 2)', \quad \mu_9 = (1\ \ -2)'.$$
$$\Sigma_1 = \begin{pmatrix} 0.3 & 0.05 \\ 0.05 & 0.3 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.2 \end{pmatrix},$$
$$\Sigma_{4,5,6,7,8} = \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}, \Sigma_9 = \begin{pmatrix} 0.3 & -0.1 \\ -0.1 & 0.3 \end{pmatrix}, \quad p_i = 1/9, \quad 1 \leq i \leq 9.$$

Last two models represent settings with a high amount of the unobserved information.

Sample sizes. We chose sample sizes 250, 500 and 1000 for Models 1, 2 and 3. Model 4 and Model 5 were constructed to have a high amount of unobserved information, so we tested them only for the sample size 5000.

Choice of the penalty weight. The theory doesn't say anything concrete about the optimal choice of the penalty weights a_n in practice. We determined the best choice for each model and each sample size by a grid search over an equidistant grid of 200 values in $[0, \frac{5}{\sqrt{n}}]$. For each value of a_n on the grid, each model and each sample size ($n \in \{250, 500, 1000\}$) Newton's method was applied 1000 times to a randomly generated sample and the number of failures n_f was counted. The grid value with the lowest n_f was used in the simulation study. It came out that penalization could not effectively reduce the number of failures. In several situations the best choice for the penalization weight was 0 (no penalization). We will discuss this point below.

Simulation results. The complete set of tables and figures can be obtained on inquiry to the author. Here only a few of them are presented. As we can see from the results, our algorithm was in many cases faster than the EM algorithm. The differences in BIC were in the most cases not significant, as the p-values from the corresponding Wilcoxon tests suggest. An exception was the SEM algorithm for Models 2 and 3 - in the cases where SEM converged, it achieved a significantly better BIC values than the rest. The only problem was, that such cases were quite rare (see Table 1.1), so we could achieve the same effect for any other algorithm by considering e.g. only the best 30% of results.

The results differ depending on the model and the sample size. For sample sizes 250 and 500 the EM algorithm usually had a better performance, followed by Newton's method and the SEM algorithm. The comparison of both the EM implementations shows an advantage for the package Mclust in most cases. In many cases when the Mclust implementation of EM was faster, the Rmixmod's one was slower than Newton's method. The constellations where the Rmixmod EM implementation was faster than the Mclust's one, were Model 3 and $n = 1000$ and Model 4 and Model 5 and $n = 5000$.

Newton's method outperformed the rest clearly for $n = 1000$ and Model 2 and $n = 5000$ and Model 4 and Model 5 - the constellations with the highest amount of the unobserved information in the EM setting.

The plots of the ecdfs of the time samples show that the ecdf of Newton's method often lies under the corresponding ecdf of the EM algorithm for small values on the x axis, and above it for higher values. That means that the EM time distribution has more mass at small values but also more mass at high values than Newton's method.

The failure counts are presented in Table 1.1. We note that the numbers for Newton's method are higher than for both EM implementations. The reason for

such a behaviour, is the quite small convergence radius of Newton's method. The penalization could not reduce the number of the failures of Newton's method (see Table 1.1). The reason for this is that most failures of Newton's method correspond to saddle points and other critical points of the log-likelihood and not to the boundary of the parameter space. One astonishing observation is that penalized Newton's method was in some cases superior over the non-penalized version, see e.g. Table 1.3 or supplement material. In fact the ecdf's of the time samples confirm this claim. A possible explanation of this fact may be that penalizing makes the log-likelihood more smoothly.

A notable observation is that the SEM algorithm was outperformed by other algorithms throughout almost all models and all sample sizes and failed to converge conspicuously often, especially in the case of Model 2 and sample sizes 500 and 1000 and Model 3 and all sample sizes (see Table 1.1). SEM was very unstable as well, as the standard deviations of the corresponding BIC samples suggest. It is quite an unexpected result, since the SEM algorithm was designed as an improvement of the EM algorithm. Gaussian mixture models seem not to be the application where the advantages of SEM justify its usage, like mixtures of distributions outside the exponential family (see e.g. [25]).

The parallel version of Newton's method was 2-6 times faster than the non-parallel one, depending on model and sample size.

1.2.4 Conclusion

The numerical experiments show that the combination of the EM algorithm and Newton's method, is in many cases faster than the pure EM algorithm. It is well known that the EM algorithm has difficulties if the fraction of missing data is high. This fraction increases with K and n, and indeed we see a clear advantage of our approach for Model 2 ($K = 5$) and sample size $n = 1000$ and Models 4 ($K = 7$) and 5 ($K = 9$) and sample size $n = 5000$. These constellations correspond to the highest amount of unobserved information in the EM setting among the tested models.

Such results would be impossible without the chosen parameterization of the covariance matrices of the components. It avoids the numerically unstable and costly matrix inversions.

However, compared to the EM algorithm, Newton's method requires much more storage and much more floating point operations per iteration. The size of the Hessian is $\mathcal{O}(K^2d^4)$, so one would guess the EM algorithm should outperform Newton's method in higher dimensions. Indeed, our implementation of Newton's

Table 1.1: Failure counts (out of 1000).

Algo	Model 1			Model 2			Model 3			Model 4	Model 5
n =	250	500	1000	250	500	1000	250	500	1000	5000	5000
NM	3	2	2	7	11	30	2	9	20	48	70
NMP	2	2	2	7	11	30	6	10	23	48	70
EMC	0	0	0	11	5	2	0	0	1	0	23
EMIX	0	0	0	6	4	0	7	0	3	0	3
SEM	52	1	0	814	637	333	397	347	222	91	249

method became slower than the EM algorithm on average machines for dimensions $d \geq 5$ (see results of Model 3).

Better implementations and faster computers should redress this problem. Quasi-Newton methods, which do not require the computation of the Hessian matrix, would redress the dimensionality issue as well. However the local quadratic convergence rate would become lost in that case.

The advantages of Newton's method should carry more weight in other mixture models, such as mixtures of t-distributions, since no explicit update formulas for all parameters in the EM algorithm exist there. Also in other settings with a high fraction of the unobserved information, where the EM algorithm is applied for the parameter estimation, Newton's method should perform faster.

One of the most relevant drawbacks of Newton's method in practice is the necessity of providing rather accurate starting values. The preceding EM iterations can resolve this problem only partly, since it is not clear a-priori how long to iterate before starting the Newton's iterations. Our approach to iterate the EM algorithm until the relative log-likelihood change fell below 1e-6 worked well, but in some few cases it was not enough to achieve the convergence region of Newton's method and algorithm failed, see Table 1.1. Xu and Jordan [49] found a representation of the EM iteration as $\theta_{k+1} = \theta_k + P_k \nabla l(\theta_k)$, where a P_k is a well-conditioned matrix, which takes the place of the negative inverse of the Hessian $-H_k^{-1}$ in NM iterations. Hence EM can be considered as a variant of the Quasi-Newton methods. A possible approach for improvement of the both methods should be the use of a convex combination of the both matrices $\omega_k P_k - (1 - \omega_k) H_k^{-1}$ as the iteration matrix. In doing so, one should adapt $\omega_k \in [0, 1]$ during the iterations. At the beginning ω_k should be near 1 and at the end near 0. The difficulty is to find appropriate criteria for adapting ω_k, it may depend on the condition number of the resulting matrix and/or on the negative definiteness of H_k.

Table 1.2: **Model 2.** Benchmarks bench(A, B)

A \ B	NM	NMP	EMC	EMIX	SEM
n = 1000					
NM	-	0	1	1	0
NMP	0	-	1	1	0
EMC	0	0	-	0	0
EMIX	0	0	1	-	0
SEM	0	0	0	0	-

Table 1.3: **Model 3.** Benchmarks bench(A, B)

A \ B	NM	NMP	EMC	EMIX	SEM
n = 250					
NM	-	0	0	-1	-1
NMP	1	-	0	-1	0
EMC	1	1	-	0	0
EMIX	1	1	0	-	0
SEM	1	0	0	0	-
n = 500					
NM	-	0	0	-1	0
NMP	1	-	0	-1	0
EMC	1	1	-	1	0
EMIX	1	1	0	-	0
SEM	0	0	0	0	-
n = 1000					
NM	-	0	0	0	0
NMP	1	-	0	0	0
EMC	1	1	-	0	0
EMIX	1	1	1	-	0
SEM	0	0	0	0	-

Table 1.4: **Model 4.** Benchmarks bench(A, B)

A \ B	NM	NMP	EMC	EMIX	SEM
			$n = 5000$		
NM	-	0	1	1	0
NMP	0	-	1	1	0
EMC	0	0	-	0	0
EMIX	0	0	1	-	0
SEM	0	0	0	0	-

Table 1.5: **Model 5.** Benchmarks *bench*(A, B)

A \ B	NM	NMP	EMC	EMIX	SEM
			$n = 5000$		
NM	-	0	1	0	0
NMP	0	-	1	0	0
EMC	0	0	-	-1	-1
EMIX	0	0	1	-	-1
SEM	0	0	1	1	-

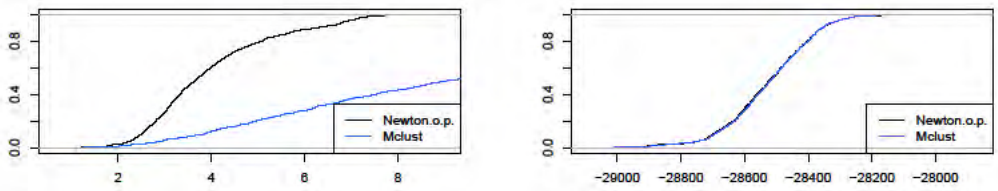

Figure 1.1: ecdf's of the time samples (left) and BIC samples (right) of Newton's method (black) and the Mclust-EM algorithm (blue) for Model 4 and sample size of fitted data 5000.

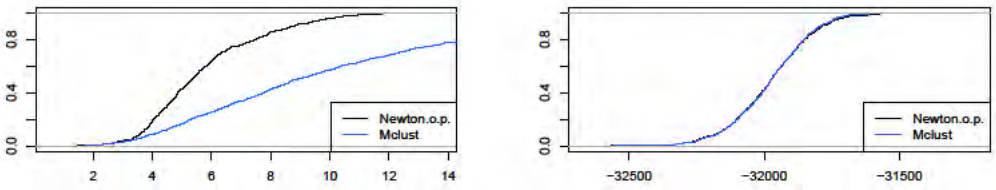

Figure 1.2: ecdf's of the time samples (left) and the BIC samples (right) of the Newton's method (black) and the Mclust-EM algorithm (blue) for Model 5 and sample size of fitted data 5000.

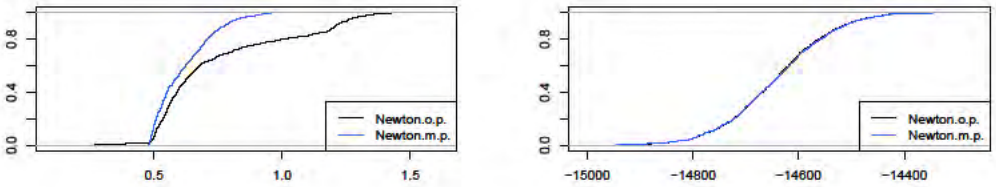

Figure 1.3: ecdf's of the time samples (left) and BIC samples (right) of Newton's method (black) and the penalized Newton's method (blue) for Model 3 and sample size of fitted data 1000.

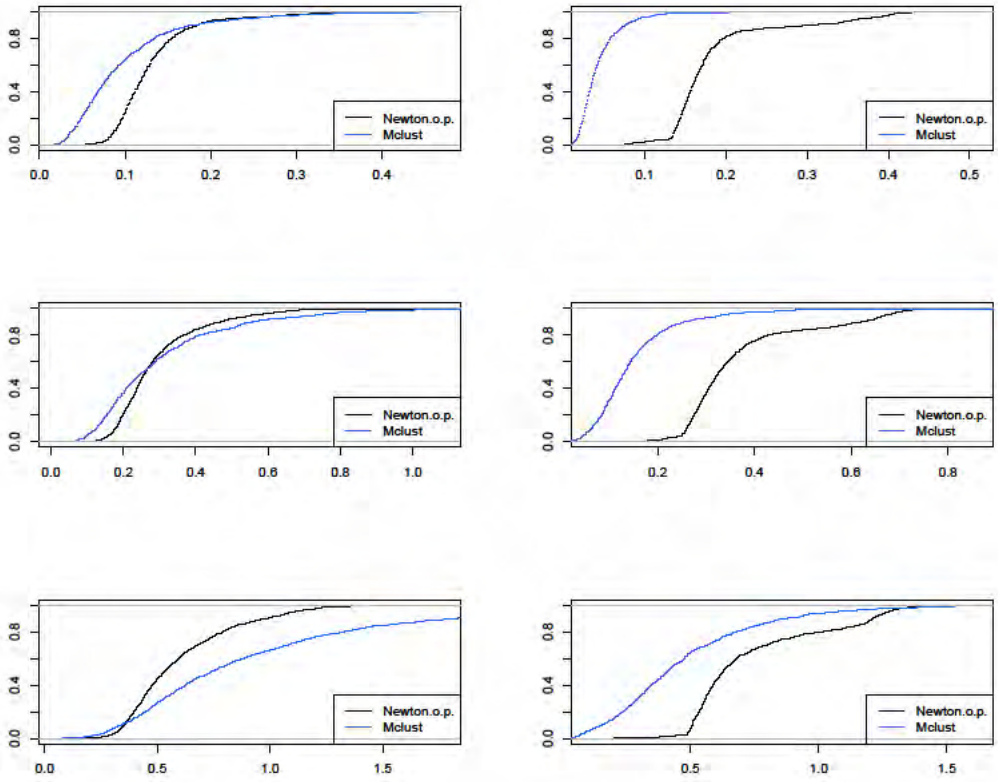

Figure 1.4: ecdf's of the time samples of Newton's method (black) and the Mclust-EM algorithm (blue) for the Models 2 (left column) and 3 (right column) and sample sizes of fitted data 250, 500, 1000 (top-down).

1.2.5 Derivatives and technical details

In the following part first and second derivatives in a form suitable for implementation, as well as some additional results are presented.

For the calculation of the derivatives we express the parameter θ as a vector in $\mathbb{R}^{K(d+\frac{1}{2}(d+1)d+1)-1}$. In order to vectorize a lower triangular matrix L, we need the following definition:

Definition 1.2.1 Let $L \in \mathbb{R}_{lt}^{d\times d}$ (a $d \times d$ lower triangular or symmetric matrix). The bijective mapping

$$\text{vech} : \mathbb{R}_{lt}^{d\times d} \to \mathbb{R}^{\frac{d(d+1)}{2}},\ L \mapsto \vec{L},$$

where $\vec{L}^i := L^{\vec{z}_i,\vec{s}_i}$ for $1 \le i \le d(d+1)/2$, with

$$\vec{z}_i := \lceil \frac{1}{2} + \sqrt{\frac{1}{4} + 2i} - 1 \rceil, \tag{1.7}$$

$$\vec{s}_i := i - \frac{\vec{z}_i(\vec{z}_i - 1)}{2}. \tag{1.8}$$

is the *half-vectorization* of the matrix L.

The mapping vech concatenates the elements of L row-wise into a vector. In each row only elements up to the diagonal are taken: $\vec{L} = (L^{i,j} : i = 1,\dots,d, j = 1,\dots,i)$.
The equations (1.7) and (1.8) have the following motivation: the coordinates of the i'th element of $\vec{L}$ in L, namely $\vec{z}_i$ (row index) and $\vec{s}_i$ (column index), must satisfy

$$i = \frac{(\vec{z}_i - 1)\vec{z}_i}{2} + \vec{s}_i,$$
$$0 < \vec{s}_i \le \vec{z}_i.$$

The only solution of this problem is given by the displayed equations. In the following we will use $\vec{L}$ as well as L in our formulas, depending on what is more suitable in a concrete situation. See also the list of notations to avoid confusion.

Proposition 1.2.2 *Let L be a $d \times d$ lower triangular or symmetric matrix, then for $1 \le j \le i \le d$ we have $L^{i,j} = \vec{L}^{k^\Delta(i,j)}$, where $k^\Delta(i,j) = \frac{i(i-1)}{2} + j$.*

In order to select the i'th row of L in $\vec{L}$ we need to pass the first $i-1$ rows, which form the first $\frac{i(i-1)}{2}$ elements in $\vec{L}$.

In the following subsection we calculate the derivatives of the log-likelihood function with respect to the parameter vector

$$\theta = (\mu_1, \ldots, \mu_K, \vec{L}_1, \ldots, \vec{L}_K, q_1, \ldots, q_{K-1}).$$

First derivatives of the mixture density. Now, the first partial derivatives of g w.r.t. the parameters μ_i, $\vec{L}_i$ and q_i will be calculated.

For $1 \leq i \leq K$ holds $\frac{\partial g}{\partial \mu_i} = p_i \phi(Y; \mu_i, L_i)(Y - \mu_i)' \Sigma_i^{-1}$. For $1 \leq i \leq K-1$ and $p = (p_1, \ldots, p_{K-1}) \in \mathbb{R}^{1 \times K-1}$ holds $\frac{\partial g}{\partial q_i} = \frac{\partial g}{\partial p} \frac{\partial p}{\partial q_i}$, where

$$\frac{\partial g}{\partial p} = (\phi_1 - \phi_K, \ldots, \phi_{K-1} - \phi_K) \qquad \text{and} \tag{1.9}$$

$$\frac{\partial p}{\partial q_i} = -\frac{2q_i}{(\sum_{j=1}^{K-1} q_j^2 + 1)^2} (q_1^2, \ldots, \overbrace{-\sum_{j=1,\, j\neq i}^{K-1} q_j^2 - 1}^{i^{th} \text{ component}}, \ldots, q_{K-1}^2)', \tag{1.10}$$

where $\phi_j = \phi(Y; \mu_j, L_j)$.
Before formulas for derivatives with respect to the covariance parameters can be obtained, we need the following properties of a lower triangular matrix L:

1. $|L| = \prod_{i=1}^{d} L^{i,i}$, and for $i \geq j$: $\frac{\partial |L|}{\partial L^{i,j}} = \begin{cases} 0 & i \neq j \\ \prod_{k=1,\, k\neq i}^{d} L^{k,k} & \text{otherwise} \end{cases}$,

2. $$\frac{\partial (Y-\mu)' L L' (Y-\mu)}{\partial L} = 2(Y-\mu)(Y-\mu)' L.$$

The notation $\frac{\partial f}{\partial L}$ for a function f which maps a matrix L onto a real number $f(L)$ means a matrix of derivatives $(\frac{\partial f}{\partial L_{i,j}})_{i,j}$. With the above formulas we obtain: for $1 \leq i \leq K$ holds

$$\frac{\partial g}{\partial \vec{L}_i} = \text{vech}\left(\frac{p_i}{|L_i|} \phi(Y; \mu_i, L_i) \left[\text{diag}(\prod_{k\neq 1}^{d} L_i^{k,k}, \ldots, \prod_{k\neq d}^{d} L_i^{k,k}) - |L_i| (Y-\mu_i)(Y-\mu_i)' L_i \right] \right).$$

Since L_i is a lower triangular matrix, we can speedup the calculations of $(q_{r,s})_{r,s} := (Y-\mu_i)(Y-\mu_i)'L_i$ by setting $M_i := (Y-\mu_i)(Y-\mu_i)'$ and considering

$$q_{r,s} = \sum_{k=1}^{d} M_i^{k,r} L_i^{k,s} = \sum_{k=s}^{d} M_i^{k,r} L_i^{k,s}.$$

Now we can calculate the gradient of the mixture-density w.r.t. θ:

$$\nabla_\theta g = \Big(\begin{array}{cccccccc} \frac{\partial g}{\partial \mu_1} & \cdots & \frac{\partial g}{\partial \mu_K} & \frac{\partial g}{\partial \vec{L}_1} & \cdots & \frac{\partial g}{\partial \vec{L}_K} & \frac{\partial g}{\partial q_1} & \cdots \quad \frac{\partial g}{\partial q_{K-1}} \end{array} \Big)'.$$

First derivatives of the log-likelihood. Derivatives of the log-likelihood function $l(\theta) = \log(\prod_{t=1}^{n} g(Y_t;\theta)) = \sum_{t=1}^{n} \log(g(Y_t;\theta))$ are obtained from the relationship $\nabla l(\theta) = \sum_{t=1}^{n} \frac{1}{g(Y_t;\theta)} \nabla_\theta g(Y_t;\theta)$.

Second derivatives of the mixture density. For Newton's method also the second derivatives of the log-likelihood function, e.g. its Hessian are required. In the first step we calculate the Hessian of the mixture density $\nabla^2_\theta g(Y,\theta)$ for a fixed Y. For two natural numbers a, b the value $\delta_a(b)$ is 1 if $a = b$ and 0 otherwise.

For the following computations we set $M_i = (Y-\mu_i)(Y-\mu_i)'$.

For $1 \le i < k \le K$ holds $\frac{\partial^2 g}{\partial q_i \partial q_k} = \frac{\partial g}{\partial p} \frac{\partial^2 p}{\partial q_i \partial q_k}$, where for $i \neq k$, $1 \le l \le K-1$:

$$\frac{\partial^2 p_l}{\partial q_i \partial q_k} = \frac{1}{(\sum_{j=1}^{K-1} q_j^2 + 1)^3} \cdot \begin{cases} 8 q_i q_k q_l^2 & l \neq i,k, \\ 4 q_i q_k ((\sum_{j=1}^{K-1} q_j^2 + 1) - 2(\sum_{j\neq i}^{K-1} q_j^2 + 1)) & l = i, \\ 4 q_i q_k (2q_k^2 - (\sum_{j=1}^{K-1} q_j^2 + 1)) & l = k. \end{cases}$$

$$\frac{\partial^2 p_l}{\partial q_i^2} = \frac{1}{(\sum_{j=1}^{K-1} q_j^2 + 1)^3} \cdot \begin{cases} 2q_l^2 \big[4q_i^2 - (\sum_{j=1}^{K-1} q_j^2 + 1)\big] & l \neq i, \\ 2(\sum_{j\neq i}^{K-1} q_j^2 + 1)\big[(\sum_{j=1}^{K-1} q_j^2 + 1) - 4q_i^2\big] & \text{else.} \end{cases}$$

$$\frac{\partial^2 g}{\partial q_i \partial \vec{L}_j} = \frac{\partial^2 g}{\partial \vec{L}_j \partial p} \frac{\partial p}{\partial q_i}, \quad \text{where for } 1 \le l \le K-1$$

$$\frac{\partial^2 g}{\partial \vec{L}_j \partial p_l} = (\delta_j(l) + \delta_j(K)) \frac{(-1)^{\delta_j(K)}}{|\vec{L}_j|} \phi(Y; \mu_j, L_j) \left[\text{vech}(\text{diag}(\prod_{k \neq 1}^{d} L_j^{k,k}, \ldots, \prod_{k \neq D}^{d} L_j^{k,k}) - |L_j| M_j \vec{L}_j) \right] \text{ and } \frac{\partial p}{\partial q_i} \text{ is given by (1.2.5).}$$

For $1 \leq i, j \leq K$ holds

$$\frac{\partial^2 g}{\partial \mu_i \partial \mu_j'} = \delta_i(j) p_i \phi(Y; \mu_i, L_i) \left[\Sigma_i^{-1} (Y - \mu_i)(Y - \mu_i)' \Sigma_i^{-1} - \Sigma_i^{-1} \right],$$
$$\frac{\partial^2 g}{\partial \vec{L}_i \partial \mu_j} = \delta_i(j) \left(\begin{array}{ccc} \frac{\partial^2 g}{\partial \vec{L}_i^1 \partial \mu_i} & \cdots & \frac{\partial^2 g}{\partial \vec{L}_i^{d(d+1)/2} \partial \mu_i} \end{array} \right),$$

where

$$\frac{\partial^2 g}{\partial \vec{L}_i^j \partial \mu_i} = \frac{p_i}{|L_i|} \phi(Y; \mu_i, L_i) \left[(Y - \mu_i)' \Sigma_i^{-1} \left(\delta_{\vec{z}_j}(\vec{s}_j) \prod_{k \neq \vec{z}_j}^{d} L_i^{k,k} - |L_i| [M_i L_i]^{\vec{z}_j, \vec{s}_j} \right) + |L_i| L_i^{\cdot, \vec{s}_j\,\prime} (Y - \mu_i)^{\vec{z}_j} + e_{\vec{z}_j} L_i^{\cdot, \vec{s}_j\,\prime} (Y - \mu_i) \right].$$

Further $\frac{\partial^2 g}{\partial q_i \partial \mu_j} = \frac{\partial^2 g}{\partial \mu_j \partial p} \frac{\partial p}{\partial q_i}$, where $\frac{\partial^2 g}{\partial p_l \partial \mu_j} = (\delta_j(l) + \delta_j(K))(-1)^{\delta_K(j)} \phi(Y; \mu_j, L_j)(Y - \mu_j)' \Sigma_j^{-1}$ and $\frac{\partial^2 g}{\partial \vec{L}_i \partial \vec{L}_j} = \delta_i(j) \left(\begin{array}{ccc} \frac{\partial^2 g}{\partial \vec{L}_i^1 \partial \vec{L}_i} & \cdots & \frac{\partial^2 g}{\partial \vec{L}_i^{d(d+1)/2} \partial \vec{L}_i} \end{array} \right)$, where

$$\frac{\partial^2 g}{\partial \vec{L}_i^j \partial \vec{L}_i} = \frac{p_i \phi(Y; \mu_i, L_i)}{|L_i|} \left[(\vec{M_i L_i}) \left(|L_i| (M_i L_i)^{\vec{z}_j, \vec{s}_j} - \delta_{\vec{z}_j}(\vec{s}_j) \prod_{k \neq \vec{z}_j}^{d} L_i^{k,k} \right) + \delta_{\vec{z}_j}(\vec{s}_j) \text{vech}(\text{diag}(\prod_{k \neq 1, \vec{z}_j}^{d} L_i^{k,k}, \ldots, \underset{j\uparrow}{0}, \ldots, \prod_{k \neq D, \vec{z}_j}^{d} L_i^{k,k})) - \xi_j \right],$$

where ξ_j is a $\frac{d(d+1)}{2}$-vector with

$$\xi_{j,p} := \delta_{\vec{z}_p}(\vec{s}_p) \prod_{k \neq \vec{z}_p}^{d} L_i^{k,k} (M_i L_i)^{\vec{z}_j, \vec{s}_j} + \delta_{\vec{z}_j}(\vec{z}_p) |L| (Y - \mu_i)^{\vec{z}_j} (Y - \mu_i)^{\vec{z}_p}.$$

Second derivatives of the log-likelihood. Next proposition is needed to obtain a formula for the Hessian of the log-likelihood.

Proposition 1.2.3 *Let $f : \mathbb{R}^d \to \mathbb{R}^d$ and $h : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable functions. Then $J_x(hf) = hJ_x(f) + f\nabla_x h'$,*

where $J_x(f)$ is the Jacobian of f w.r.t. x and $\nabla_x h$ is the gradient of h w.r.t. x.

Proof. $hf(x) = \begin{pmatrix} h(x)f_1(x) \\ \vdots \\ h(x)f_D(x) \end{pmatrix}$ and using the product rule we obtain

$$\nabla_x(hf_j) = h\nabla_x f_j + f_j \nabla_x h.$$

□

The goal now is to calculate the Jacobian of $\nabla_\theta l(\theta) = \sum_{t=1}^{n} \frac{1}{g(Y_t;\theta)} \nabla_\theta g(Y_t;\theta)$. Proposition 1.2.3 with $f = \nabla_\theta g$ and $h = \frac{1}{g}$ yields that $\nabla_\theta \frac{1}{g(Y_t;\theta)} = \frac{-1}{g(Y_t;\theta)^2} \nabla_\theta g(Y_t;\theta)$. Summing over all t's finally yields

$$\nabla_\theta^2 l = \sum_{t=1}^{n} \frac{1}{g(Y_t;\theta)} \left(\nabla_\theta^2 g(Y_t;\theta) - \frac{1}{g(Y_t;\theta)} \nabla_\theta g(Y_t;\theta) \nabla_\theta' g(Y_t;\theta) \right).$$

Derivatives of the penalty terms. The penalty function used in our algorithm is given by

$$s_n(L_1, \ldots, L_K) = -a_n \sum_{i=1}^{K} \big(\mathrm{tr}(S L_i {L_i}') + \log \frac{1}{|L_i|^2} \big).$$

Now we omit the index of the covariance parameter. Hence L represents any of the $L_1, \ldots, L_K$. For $1 \le s \le \frac{d(d+1)}{2}$ holds

$$\frac{\partial s_n}{\partial \vec{L}^s} = -a_n \big(2 \sum_{k=\vec{s}_s}^{d} S^{k,\vec{z}_s} L^{k,\vec{s}_s} - \frac{2\delta_{\vec{z}_s}(\vec{s}_s)}{\vec{L}^s} \big),$$

$$\frac{\partial^2 s_n}{\partial L_s^{\Delta} \partial L_t^{\Delta}} = -a_n \big(2\delta_{\vec{s}_s}(\vec{s}_t) S^{\delta_{\vec{z}_t}(\vec{s}_t)} + \frac{2\delta_{\vec{z}_s}(\vec{s}_s)\delta_s(t)}{(\vec{L}^s)^2} \big).$$

Derivatives of the parameter transformation ψ. We calculate the derivatives of

$$\psi(\mu_1, \ldots, \mu_K, L_1, \ldots, L_K, q_1, \ldots, q_{K-1}) = (\mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K, p_1, \ldots, p_{K-1}).$$

The derivatives $\frac{\partial\psi}{\partial q}$ are given in (1.10). Now we calculate the partial derivatives of ψ w.r.t. L_i. We represent the matrices as vectors and calculate the derivatives of the function $\vec{L} \mapsto \text{vech}(LL'^{-1})$, for a lower triangular matrix L. This function can be expressed as a composition $\psi_1 \circ \psi_2$ for $\psi_1 : \vec{A} \mapsto \vec{A^{-1}}$, and $\psi_2 : \vec{L} \mapsto \text{vech}(LL')$.

Proposition 1.2.4 *i) Let A be a regular matrix, $i \geq j$ two integers. Then*

$$\frac{\partial\psi_1}{\partial A_{ij}} = -vech(A^{-1}e_i{e_j}'A^{-1})$$

ii) Let L be a lower triangular Matrix, $i \geq j, p \geq q$ four integers. Then

$$\frac{\partial{\psi_2}_{p,q}}{\partial L_{ij}} = \delta_i(p)L_{q,j} + \delta_i(q)L_{p,j}$$

Proof. i)

$$A^{-1}A = I \Rightarrow \frac{\partial A^{-1}A}{\partial A_{ij}} = \frac{\partial A^{-1}}{\partial A_{ij}}A + A^{-1}\underbrace{\frac{\partial A}{\partial A_{ij}}}_{=e_i{e_j}'} = 0$$

$$\Leftrightarrow \frac{\partial A^{-1}}{\partial A_{ij}} = -A^{-1}\frac{\partial A}{\partial A_{ij}}A^{-1}$$

ii) It follows immediately from $(LL')_{i,j} = L'_{i,\cdot}L_{j,\cdot}$.

In the following section we will consider some theoretical issues behind the estimation of Gaussian mixtures via likelihood maximization.

1.3 Penalized estimation of multivariate Gaussian mixture models

Now, we consider the theoretical side of the problem of estimating the parameters of a multivariate Gaussian mixture with K components by maximizing the likelihood function. As already mentioned before, this approach has a theoretical drawback: the likelihood function is unbounded, and the interesting maxima are local maxima in the interior of the parameter space. Consider an estimator with $\hat{\mu}_1 = Y_1$, $|\hat{\Sigma}_1| = \varepsilon$, $\hat{\mu}_2 \in \mathbb{R}^d$ arbitrary, $\hat{\Sigma}_2 = I$, $\hat{p}_1 = 1/2$. Then the likelihood function tends to infinity as $\varepsilon \to 0$ and hence the MLE is not consistent.

Two basic strategies for overcoming the unboundedness were studied in the literature: restricted optimization and penalization of the likelihood. In the first case a lower bound on the variances or their ratios is imposed; see e.g. Hathaway [20]. In the second case a term which penalizes small variances or ratios of variances is added to the log-likelihood; see e.g. Ciuperca et al. [13], Tanaka [44], Chen et al. [12], Chen and Tan [11]. The second approach has some advantages over the first one - there is no tuning constant to choose and the penalty function actually disappears with increasing sample size.

In the following, we discuss the consistency proof of the penalized MLE from Chen and Tan [11]. Among the above papers on consistency of the penalized MLE it is the most interesting one in the context of Gaussian mixtures, since it treats the multivariate case. Adjusting the penalty magnitude is an important issue and requires an assessment of the number of observations with a high likelihood contribution. Such an assessment is given in Lemma 2 in Chen and Tan [11]. However its proof seems to contain a soft spot and I was not able to fix it. In Section 1.3.1 we elaborate on the soft spot in detail. In Section 1.3.2 we give an alternative proof of a similar statement based on a uniform law of iterated logarithm. This allows us to make Chen and Tan's nice consistency proof fully rigorous. The following result can be found in Alexandrovich [4].

1.3.1 Outline of consistency proof of Chen and Tan

In the following, let Θ be the set of K-component mixture parameters in a usual parameterization $(\mu_j, \Sigma_j, p_j,\ 1 \leq j \leq K)$, where $\mu \in \mathbb{R}^d,\ \Sigma \in \mathcal{P}^d,\ p \in \triangle^{K-1}$ and $\mathcal{P}^d$ is the set of $d \times d$ symmetric positive definite matrices. Two parameters are considered as equivalent iff they induce the same distribution. θ_0 denotes as usual a true parameter. The proof has roughly the following scheme:

1 Divide the parameter space Θ into $K+1$ disjoint subsets $\Gamma_1, \ldots, \Gamma_{K+1}$ where each subset is characterized by the number of components which covariances are bounded away from zero. The subset where all covariances are bounded away from zero, Γ_{K+1}, is regular and contains the true parameter θ_0 so the classical MLE theory as in Wald [48] and Kiefer and Wolfowitz [27] can be applied, see Subsection 1.3.3.

2 Show that, asymptotically, the penalized MLE $\hat{\theta}_n^{pMLE}$ a.s. does not lie in any subset except the regular one, that is

$$\sup_{\theta \in \Gamma_i} l_n(\theta) + p_n(\theta) - l_n(\theta_0) - p_n(\theta_0) \to -\infty, \quad i \in \{1, \ldots, K\},$$

almost sure, where $p_n : \Theta \to \mathbb{R}$ is a penalty function.

The second step is quite involved and will be outlined more precisely. The penalty function p_n fulfils several conditions, see Chen and Tan [11]. Recall the key condition C3: $\tilde{p}_n(\Sigma) \leq 4(\log n)^2 \log|\Sigma|$ for $|\Sigma| < cn^{-2d}$, where $p_n(\theta) = \sum_{j=1}^{K} \tilde{p}_n(\Sigma_j)$ and c some positive constant. This condition is imposed in order to rule out the damaging effect of components with degenerate covariance matrices. It will turn out that $4(\log n)^2$ is actually not sufficient.

A key element of the proof is a uniform assessment of the number of observations, with a high contribution to the likelihood. These are observations, that are located inside certain critical regions. It turns out that an appropriate choice for such critical regions is ellipses

$$\tilde{A}(\mu, \Sigma) := \{y \in \mathbb{R}^d \ : \ (y-\mu)'\Sigma^{-1}(y-\mu) \leq (\log|\Sigma|)^2\},$$

where μ and Σ correspond to a degenerate component of the point at which the likelihood is evaluated. Figure 1.5 demonstrates the idea.

The contribution of observations inside such a set will be ruled out by the penalty function and the one outside can be shown to be small enough. Precisely, following bounds are used

$$\varphi(y; \mu, \Sigma) \leq \begin{cases} |\Sigma|^{-\frac{1}{2}} & y \in \tilde{A}(\mu, \Sigma) \\ e^{-\frac{1}{4}(y-\mu)'\Sigma^{-1}(y-\mu)} & \text{otherwise.} \end{cases}$$

A statement of the form

$$H_n(\mu, \Sigma) := \sum_{i=1}^{n} \mathbf{1}_{Y_i \in \tilde{A}(\mu,\Sigma)} \leq a(n) + b(n, |\Sigma|), \tag{1.11}$$

for all $\mu \in \mathbb{R}^d$, $\Sigma \in \mathcal{P}^d$ almost sure, is needed, where $a(n) = o(n)$ and $b(n, s) = O(n)$ for each s and $b(n, s) \log s^{-1/2} \to 0$ as $s \to 0$. An important detail here is that the *almost sure* statement has to hold simultaneously for all tuples (μ, Σ) and not solely for each one. Given any statement with these properties one can prove the consistency of the penalized MLE, following arguments from Chen and Tan [11], if the penalty function fulfils a modified condition C3: $\tilde{p}_n(\Sigma) \leq a(n) \log|\Sigma|$ for $|\Sigma| \leq cn^{-2d}$.

We will consider and study the same questions more precisely in the context of penalized estimation of Gaussian hidden Markov models later in the thesis.

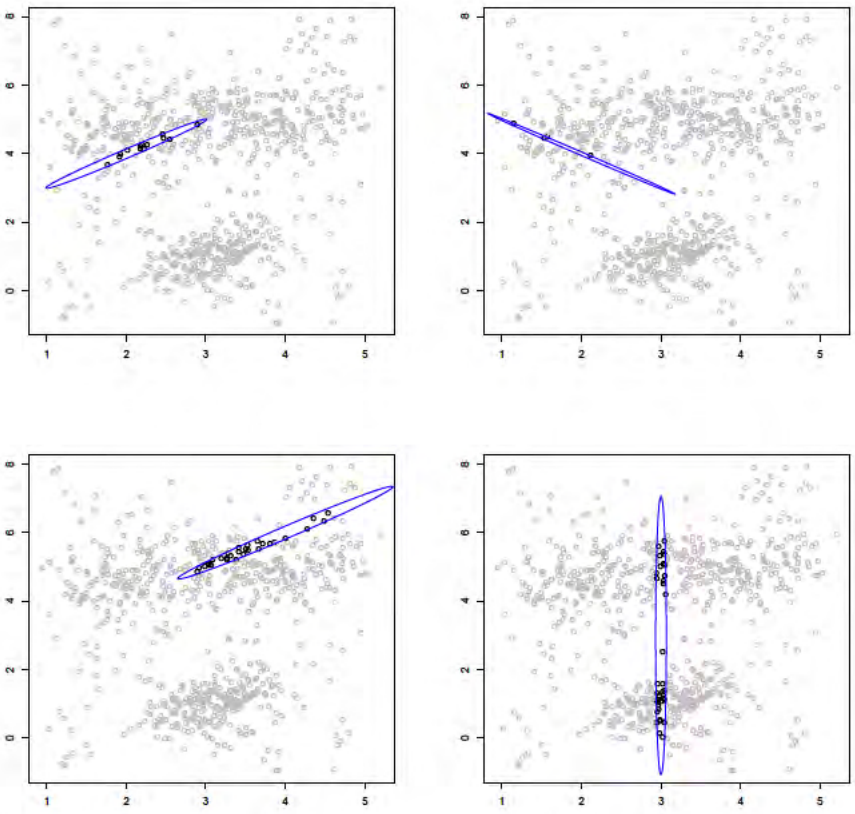

Figure 1.5: Illustration of the critical regions for the consistency proof of penalized MLE

Chen and Tan [11] claimed essentially the following bound (Lemma 2)

$$H_n(\mu, \Sigma) \leq 4(\log n)^2 + 8nM|\Sigma|^{1/2d} \log |\Sigma|, \tag{1.12}$$

for all (μ, Σ) with $|\Sigma| < \exp(-4d)$ a.s., where M is an upper bound of the true mixture density. The proof uses an ascription to the univariate case, which was proved in Chen et al. [12] by applying a Bernstein inequality and Borel-Cantelli Lemma. We omit further details of this involved proof and refer to the source. Instead, we pay our attention to the ascription, which actually does not work. The argument behind the ascription is as follows

$$\begin{aligned}
\{y \in \mathbb{R}^d \,:\, (y-\mu)'\Sigma^{-1}(y-\mu) \leq (\log|\Sigma|)^2\} & \\
&= \{y \in \mathbb{R}^d \,:\, \sum \lambda_j^{-1} |a_j'(y-\mu)|^2 \leq (\log|\Sigma|)^2\} \\
&\subseteq \{y \in \mathbb{R}^d \,:\, |a_j'(y-\mu)| \leq -\sqrt{\lambda_j}\log|\Sigma|,\ 1 \leq j \leq d\} \\
&\subseteq \{y \in \mathbb{R}^d \,:\, |a_j'(y-\mu)| \leq -\sqrt{\lambda_1}\log|\Sigma|\},
\end{aligned}$$

where $a_1, \ldots, a_d$ and $\lambda_1 \leq \ldots \leq \lambda_d$ are unit length eigenvectors and the corresponding eigenvalues of Σ respectively.

Further one argues that for every bounded set $B \subset \mathbb{R}^d$, there exists a finite subset of the unit d-sphere $Q \subset S^{d-1}$, such that for every $a \in S^{d-1}$ there exists $b \in Q$ with the following property

$$\{y \in B : |a'(y-\mu)| \leq -\sqrt{\lambda_1}\log|\Sigma|\} \subseteq \{y \in B : |b'(y-\mu)| \leq -\sqrt{2\lambda_1}\log|\Sigma|\}. \tag{1.13}$$

and concludes

$$H_n(\mu, \Sigma) \leq \max_{b \in Q} \sum_{i=1}^{n} \mathbf{1}_{\{|b'(Y_i-\mu)| \leq -\sqrt{2\lambda_1}\log|\Sigma|\}},$$

for every $\Sigma \in \mathcal{P}^d$. Hence the problem is reduced to univariate samples $b'Y_1, \ldots b'Y_n$ for finitely many $b \in S^{d-1}$. But the conclusion is not correct, since the inclusion (1.13) holds only given a fixed, bounded set B but not on the whole $\mathbb{R}^d$. I found no easy way to correct the ascription to the univariate case. However, there is an alternative, more easy approach.

1.3.2 Approach based on the uniform law of iterated logarithm

For the next statements the term of the Vapnik-Chervonenkis dimension of a class of sets is needed. This combinatorial concept serves for characterization of the complexity of a class of sets.

Definition 1.3.1 Let $\mathcal{X}$ be a complete separable metric space, $\mathcal{C} \subset 2^{\mathcal{X}}$ a family of subsets, $D \subset \mathcal{X}$ any finite subset. The *shatter coefficient* of $\mathcal{C}$ with respect to D is defined by

$$S(D : \mathcal{C}) := |\{C \cap D \ : \ C \in \mathcal{C}\}|. \tag{1.14}$$

The *VC dimension* of $\mathcal{C}$, $dim(\mathcal{C})$ is the largest integer $k \in \mathbb{N}$ such that $S(D : \mathcal{C}) = 2^k$ for some k-element subset D of $\mathcal{X}$. If for every k there exists a finite k-element subset $D \subset \mathcal{X}$ such that $S(D : \mathcal{C}) = 2^k$, then $dim(\mathcal{C}) = \infty$.

A class $\mathcal{C}$ with a finite VC dimension is called a *VC class*.

A class $\mathcal{F}$ of real valued functions $\mathcal{X} \to \mathbb{R}$ is called a *VC-graph class* if the collection of all sub-graphs of the functions in $\mathcal{F}$ forms a VC class of sets in $\mathcal{X} \times \mathbb{R}$.

VC classes have some comfortable properties, like being Glivenko-Cantelli or even Donsker classes, see e.g. van der Vaart and Wellner [47].

If $\mathcal{C}$ is a VC class, then the class $\mathcal{F} := \{1_C \ : \ C \in \mathcal{C}\}$ of indicator functions is a VC-graph class satisfying conditions of Theorem 2.13 from Alexander [3] and the next statement follows.

Theorem 1.3.2 *Let $\mathcal{C} \subset \mathcal{B}^d$ be a VC class of sets, $(Y_n)_{n\in\mathbb{N}}$ a d-dimensional i.i.d. process. Then a.s.*

$$\limsup_{n\to\infty} \sup_{C\in\mathcal{C}} \frac{|\sum_{i=1}^n 1_C(Y_i) - nP_{Y_1}(C)|}{\sqrt{2n \log\log n}} = \sup_{C\in\mathcal{C}} (P_{Y_1}(C)(1 - P_{Y_1}(C)))^{1/2}. \tag{1.15}$$

Hence we have the following corollary.

Corollary 1.3.3 *Let $(Y_n)_{n\in\mathbb{N}}$ be a d-dimensional i.i.d. process and*

$$\mathcal{E}_d := \left\{ \{y \in \mathbb{R}^d \ : \ (y-\mu)'A(y-\mu) \le 1\} \ : \quad \mu \in \mathbb{R}^d, \quad A \in \mathbb{R}^{d\times d} \ \ s.p.d. \ \right\}.$$

Then a.s. there exits $N \in \mathbb{N}$ such that

$$\sum_{i=1}^n \mathbf{1}_{\{Y_i \in C\}} \le \frac{3}{4}\sqrt{n \log\log n} + nP_{Y_1}(C) \quad \text{for all } n \ge N \text{ and all } C \in \mathcal{E}_d \tag{1.16}$$

Remark: The constant 3/4 can be replaced by any other constant greater than $\sqrt{2}/2$.

Proof. Akama and Irie [2] have shown that the VC-dimension of the set $\mathcal{E}_d$ is $(d^2+3d)/2$. From Theorem 1.3.2 follows: for any $\varepsilon > 0$ a.s. there exists $N \in \mathbb{N}$ such that

$$\sup_{C\in\mathcal{E}_d} \frac{\sum_{i=1}^n 1_C(Y_i) - nP_{Y_1}(C)}{\sqrt{2n\log\log n}} \leq \sup_{c\in\mathcal{E}_d}(P_{Y_1}(C)(1-P_{Y_1}(C))^{1/2} + \varepsilon \text{ for all } n \geq N$$
$$\Rightarrow \sum_{i=1}^n 1_C(Y_i) \leq nP_{Y_1}(C) + (1/2+\varepsilon))\sqrt{2n\log\log n} \text{ for all } n \geq N,\ C \in \mathcal{E}_d.$$

□

With the above corollary we can a.s. uniformly bound the number of i.i.d. observations generated by a bounded Lebesgue density falling into an elliptical region in $\mathbb{R}^d$.

Corollary 1.3.4 *Let $(Y_n)_{n\in\mathbb{N}}$ be i.i.d. variables with a bounded Lebesgue density f, $M := \sup_y f(y)$. Then a.s. there exists $N \in \mathbb{N}$ such that*

$$\sum_{i=1}^n \mathbf{1}_{\{(Y_i-\mu)'\Sigma^{-1}(Y_i-\mu)\leq(\log|\Sigma|)^2\}} \leq \frac{3}{4}\sqrt{n\log\log n} + \frac{nM\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}|\Sigma|^{\frac{1}{2}}|\log|\Sigma||^d \quad (1.17)$$

for every $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d\times d}$ symmetric positive definite and $n \geq N$.

Proof. First we show $P_{Y_1}(C) \leq M\frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}|\Sigma|^{\frac{1}{2}}|\log|\Sigma||^d$, where C is the ellipse $C := \{(y-\mu)'\Sigma^{-1}(y-\mu) \leq (\log|\Sigma|)^2\}$ and then we apply Corollary 1.3.3.

P_{Y_1} has Lebesgue density $f \leq M$. Hence $P_{Y_1}(C) \leq M\lambda^d(C)$. The Lebesgue measure of the ellipsoid C is given by $\lambda^d(C) = |\Sigma|^{1/2}\lambda^d(\{y'y \leq (\log|\Sigma|)^2\})$ by the invariance of λ^d w.r.t. translations and the substitution rule. For the measure of the sphere it holds $\lambda^d(\{y'y \leq (\log|\Sigma|)^2\}) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}|\log|\Sigma||^d$. □

We conclude, a bound as in (1.11) with functions $a(n) = \sqrt{2n\log\log n}$ and $b(n,|\Sigma|) = \frac{nM\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}|\Sigma|^{\frac{1}{2}}|\log|\Sigma||^d$ is obtained.

1.3.3 Wald's consistency proof

For the sake of completeness, we apply in this section Wald's classical consistency proof for the MLE as given in Ferguson [16] along with a compactification approach

as in Kiefer and Wolfowitz [27] to show that $\sup_{\theta\in\Gamma_{K+1}} l_n(\theta)+p_n(\theta)$ a.s. converges towards θ_0.

We consider the metric $d_c(\theta,\eta)=\sum_{s=1}^{r}|\arctan(\theta_s)-\arctan(\eta_s)|$, where r is the dimension of the arguments, add limits of Cauchy sequences with respect to d_c to Γ_{K+1} and denote the closure by $\bar{\Gamma}_{K+1}$. Following conditions are satisfied:

W1. $\theta\mapsto\log g(y;\theta)$ is continuous on $\bar{\Gamma}_{K+1}$ for all y.

W2. $\forall\,\theta\in\bar{\Gamma}_{K+1}\,\forall\,\rho>0$ small enough, $\sup_{d_c(\theta',\theta)\le\rho}\log g(Y;\theta')$ is measurable.

W3. $\forall\,\theta\in\bar{\Gamma}_{K+1}\,\exists\,\rho>0$, such that $\mathbb{E}_0|\sup_{d_c(\theta',\theta)\le\rho}\log g(Y;\theta')|<\infty$.

W4. The metric space $(\bar{\Gamma}_{K+1},d_c)$ is compact.

Condition W1 follows from the definition of a Gaussian density. Condition W2 follows from the continuity of $\log g(y;\cdot)$ and the fact that $\mathbb{Q}$ is dense in $\mathbb{R}$, which is why building maxima over the set $\{\theta'\,|\,d_c(\theta',\theta)\le\rho\}$ is the same as building maxima over a dense countable subset. Condition W3 is satisfied since on $\bar{\Gamma}_{K+1}$ Gaussians are bounded from above by $\frac{1}{\sqrt{\varepsilon}}$ for a fixed $\varepsilon>0$. Condition W4 is obvious, since $\bar{\Gamma}_{K+1}$ is closed and a subset of $\{\theta\,|\,d_c(0,\theta)\le\frac{r\pi}{2}\}$.

Theorem 1.3.5 *The penalized maximum likelihood estimator $\hat{\theta}_n^{pMLE}$ converges towards a true parameter θ_0 a.s.*

Proof. We already know that $\hat{\theta}_n^{pMLE}$ is a.s. located in $\bar{\Gamma}_{K+1}$. For $y\in\mathbb{R}^d$, $\theta\in\bar{\Gamma}_{K+1}$ we set $U(y,\theta):=\log g(y;\theta)-\log g(y;\theta_0)$, where $g(Y;\theta)=0$ if some entry in θ is ∞ and $\psi(y,\theta,\rho):=\sup_{d_c(\theta',\theta)\le\rho}U(y,\theta')$, where $\rho>0$. We know from W2 that functions ψ are measurable.

Furthermore, by continuity we have $\psi(y,\theta,\rho)\searrow U(y,\theta)$ as $\rho\searrow 0$ and by the Monotone Convergence Theorem

$$\mathbb{E}_0\psi(Y,\theta,\rho)\searrow\mathbb{E}_0U(Y,\theta)=-H(\theta_0,\theta)\quad\text{ as }\rho\searrow 0. \tag{1.18}$$

The negative Kullbak-Leibler divergence $-H(\theta_0,\theta)$ is negative if $\theta\not\sim\theta_0$ (θ induces another distribution than θ_0) due to the identifiability of Gaussian mixtures.

Now, let S be any closed subset of $\bar{\Gamma}_{K+1}$ not containing any of the points that induce the same distribution as θ_0 (the equivalence class of θ_0). Since $\bar{\Gamma}_{K+1}$ is compact, S is too. Let $\epsilon>0$. For every $\theta\in S$ let ρ_θ be a radius such that $\mathbb{E}_0\psi(Y,\theta,\rho_\theta)<-H(\theta_0,\theta)+\epsilon$. Since S is compact, there exists a finite cover by balls $B_1,\dots,B_m$ with radii $\rho_i=\rho_{\theta_i}$ and centers at θ_i for $i=1,\dots,m$.

By definition of ψ it holds $\frac{1}{n}\sum_{i=1}^{n} U(Y_i,\theta) \leq \frac{1}{n}\sum_{i=1}^{n}\psi(Y_i,\theta_j,\rho_j)$, where $\theta \in B_j$ and hence

$$\sup_{\theta\in S}\frac{1}{n}\sum_{i=1}^{n} U(Y_i,\theta) \leq \sup_{1\leq j\leq m}\frac{1}{n}\sum_{i=1}^{n}\psi(Y_i,\theta_j,\rho_j).$$

Now, by strong law of large numbers and (1.18) a.s. there exists $N \in \mathbb{N}$, such that $\frac{1}{n}\sup_{1\leq j\leq m}\sum_{i=1}^{n}\psi(Y_i,\theta_j,\rho_j) \leq \sup_{1\leq j\leq m} -H(\theta_0,\theta_j)+\epsilon$ and hence

$$\sup_{\theta\in S}\frac{1}{n}\sum_{i=1}^{n} U(Y_i,\theta) \leq \sup_{1\leq j\leq m} -H(\theta_0,\theta_j)+\epsilon$$

for all $n \geq N$.

The term on the right hand side is negative for ϵ small enough. For the penalized MLE a.s. holds

$$\hat{\theta}_n^{pMLE} = \sup_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n} U(Y_i,\theta) + o(1) \geq 0,$$

for n large enough, since the sum is zero for $\theta = \theta_0$ This shows, that $\hat{\theta}_n^{pMLE} \notin S$ and finishes the proof. □

1.3.4 Conclusion

The soft spot in the consistency poof in Chen and Tan [11] was identified, namely the ascription to the univariate case in Lemma 2 there. The introduced alternative in form of Corollary 1.3.4 fits almost seamless in Chen's consistency proof. Merely the condition C3 on the penalty function has to be strengthened to $\tilde{p}_n(\Sigma) \leq (\frac{3}{4}\sqrt{n\log\log n})\log|\Sigma|$ for $|\Sigma| < cn^{-2d}$ for some $c > 0$. However, it is not a problem, since the example penalty function with $\tilde{p}_n(\Sigma) = -n^{-1}(\mathrm{tr}(\Sigma^{-1}) + \log|\Sigma|)$ fulfills this requirement. To see this, assume $|\Sigma| < n^{-2d}$. Then it holds for the eigenvalues of Σ: $\prod_{i=1}^{d}\lambda_i < n^{-2d}$ and hence $\lambda_1 = \lambda_{min} < n^{-2}$. Now, write the trace as the sum of the eigenvalues: $\mathrm{tr}(\Sigma^{-1}) = \lambda_1^{-1} + \ldots + \lambda_d^{-1} > \lambda_{min}^{-1} > n^2$. Finally $-n^{-1}(\mathrm{tr}(\Sigma^{-1}) + \log|\Sigma|) < -n + n^{-1}2d\log n < -(\frac{3}{4}\sqrt{n\log\log n})2d\log n$ for n large enough.

The theoretical background for the new approach is given by Alexander's uniform law of iterated logarithm for VC classes. Elaborated arguments involving Bernstein's inequality and Borel-Cantelli lemma needed for the one-dimensional case as in Chen et al. [12] are avoided and the proof becomes thereby shorter and more simple.

Moreover, the introduced approach, together with the general proof principle as in Chen and Tan [11] resp. Chen et al. [12] can be used to prove consistency results for penalized MLE for mixtures of distributions with similar properties, like gamma distributions.

Once, the penalized MLE is shown to lie in a regular subset of the parameter space, Wald's consistency proof along with a compactification argument from Kiefer and Wolfowitz [27] applies straightforward.

2 Penalized estimation of Gaussian hidden Markov models

Hidden Markov models build a wide class of general-purpose models for describing weakly dependent stochastic processes and can be regarded as a generalization of finite mixtures models.

A hidden Markov model with $K \in \mathbb{N}$ states is a bivariate stochastic process $(X_t, Y_t)_{t\in\mathbb{N}}$ such that $(Y_t)_{t\in\mathbb{N}}$ are independent given $(X_t)_{t\in\mathbb{N}}$, $X_t \in \{1, \dots, K\}$ and $X_t \mid X_1^{t-1} = X_t \mid X_{t-1}$, and $Y_t \mid X_1^t = Y_t \mid X_t$. By the notation Y_m^n for $n > m$ we denote the vector $(Y_m, Y_{m+1}, \dots, Y_n)'$.

The process (X_t) is a first order Markov chain and will be referred to as the state process. If the probabilities $\mathbb{P}(X_t = i \mid X_{t-1} = j)$ do not depend on t, the Markov chain is called *homogeneous*. A Markov chain is called *irreducible* iff the corresponding graph is irreducible, that is there exists a path between every two vertexes. A Markov chain is called *stationary* iff for every finite integer tuple $(t_1, \dots, t_k)$ and any $g \in \mathbb{N}$ the equality $(X_{t_1}^{t_k}) \stackrel{d}{=} (X_{t_1+g}^{t_k+g})$ holds. An irreducible (homogeneous, discrete-time, finite state space) Markov chain with t.p.m. Φ has a unique, strictly positive stationary distribution π, i.e. $\pi' = \pi'\Phi$ and $\pi > 0$ componentwise, see e.g. Zucchini and MacDonald [50].

A Markov chain is called *aperiodic* if $\gcd\{t > 1 \mid p_{ii}^{(t)} > 0\} = 1$, where $p_{ii}^{(t)} = \mathbb{P}(X_t = i \mid X_1 = i)$ for every i. In words aperiodicity means, that there is no deterministic structure in the set of the returning times - the fact of being in state i at time 1 does not exclude the possibility of being in this state at an arbitrarily other time in the future. Aperiodicity and irreducibility is equivalent to primitivity of the transition matrix, meaning, that it has only one simple eigenvalue on the complex unit circle. For an irreducible non-negative matrix it is sufficient to have at least one non-zero element on the diagonal in order to be primitve. A good reference on Markov chains is Norris [37].

In the following only homogeneous, irreducible and aperiodic HMMs will be considered.

The state process cannot be observed (is hidden) and all inference has to be based on the observations of (Y_t). Such situations occur when the distribution of (Y_t) is determined by the value of an underlying group membership Markov process (X_t). There are many application areas for HMMs such as speech-, face-, handwriting recognition, biological sequence analysis, earthquakes prediction, finance etc., see e.g. Zucchini and MacDonald [50], Rabiner and Juang [41].

Often the state-dependent distributions $Y_t \mid X_t = k$ are determined by a finite-dimensional euclidean parameter, like in the case of Gaussian HMMs. Then the law of the process (Y_t, X_t) is determined by the t.p.m. and the vector of state-dependent parameters.

An important task in the context of HMMs is estimation of the underlying parameter, which is often solved by maximizing the log-likelihood function. In the case of Gaussian HMMs however, like in the case of Gaussian mixtures, a direct maximization has a theoretical drawback since the objective function is unbounded. Consider a two-state HMM and an estimator with $\hat{\mu}_1 = Y_1$, $\hat{\sigma}_1 = \varepsilon$, $\hat{\mu}_2 \in \mathbb{R}$ arbitrary, $\hat{\sigma}_2 = 1$, $\hat{\Phi}$ irreducible. Then the likelihood function tends to infinity as $\varepsilon \to 0$ and hence the MLE is not consistent. The multivariate i.i.d. case was treated in Chapter 1, Section 1.3.

Although the unboundedness has no serious impact on the practice, since maximization algorithms, like EM, search for local maxima and converge only seldom against degenerate solutions, it should be desirable to eliminate this theoretical drawback by introducing a consistent estimator.

The state-dependent parameters of a HMM can be consistently estimated by maximizing the marginal mixture log-likelihood, or equivalently the HMM likelihood under a independence assumption (IMLE), under some technical conditions, see Lindgren [31] and references therein. One necessary condition is $\lim_{\theta \to \partial\Theta} \varphi(y; \theta) = 0$ except on a zero-measure set, independent of the limit of θ, where $\varphi(\cdot; \theta)$ is the state-dependent density. This condition is violated in our case as indicated above.

In the following section, a two-stage procedure is proposed for a consistent estimation of the parameters of a Gaussian HMM. In the first stage, the parameters of the marginal distribution of the observed process are estimated by maximizing a penalized mixture likelihood. Some ideas from Chen et al. [12] are used, where consistency of a penalized MLE for Gaussian mixtures is shown. The main difficulty during the generalization of that result is a more complicated large deviations behaviour of HMM samples.

In the second stage, the full HMM likelihood is maximized over a neighbourhood of the estimates from the stage 1. Since this neighbourhood is regular and contains the true parameter of the HMM for n large enough, the consistency result from Leroux [30] can be applied. The maximization in each step can be done with the EM algorithms for Gaussian mixture models and for HMMs respectively.

2.1 The model and main results

In what follows θ_0 denotes a true parameter of the HMM, θ_0^{mix} a true parameter of the marginal mixture and F the true marginal distribution function. Y_1^n is as before a shorthand for $(Y_1, \ldots, Y_n)$.

The matrix Φ_0 is assumed to be aperiodic and irreducible. In this chapter we let the hidden Markov model start at $-\infty$, so that it can be assumed stationary. This approach is sensible, since the initial distribution is not subject of the estimation and has no influence on the asymptotic properties of the log-likelihood.

Definition 2.1.1 Let $(X_t, Y_t)_{t\in\mathbb{Z}}$ be a stochastic process, where $(Y_t)_{t\in\mathbb{Z}}$ are independent given $(X_t)_{t\in\mathbb{Z}}$, which is a homogeneous first order Markov chain. Furthermore

$$X_t \in \{1, \ldots, K\}, \tag{2.1}$$

$$Y_t \mid (X_s)_{s\in\mathbb{Z}} \stackrel{d}{=} Y_t \mid X_t, \tag{2.2}$$

$$Y_t \mid X_t = k \stackrel{d}{=} N(\mu_{0,k}, \sigma_{0,k}^2). \tag{2.3}$$

The process $(X_t, Y_t)_{t\in\mathbb{Z}}$ is called a *Gaussian hidden Markov model*. In the special case where $(X_t)_{t\in\mathbb{Z}}$ are independent, the process $(X_t, Y_t)_{t\in\mathbb{Z}}$ corresponds to a finite Gaussian mixture model as defined in Chapter 1.

The set of possible HMM parameters will be denoted by

$$\Theta^{full} = \{(\mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2, \Phi) \mid \mu_j \in \mathbb{R}, \quad \sigma_j^2 \in (0, \infty),\ j = 1 \ldots K, \quad \Phi \in \mathcal{T}\}.$$

The set of possible parameters of a Gaussian mixture for the first stage of the algorithm will be denoted by

$$\Theta^{mix} = \{(\mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2, \pi) \mid \quad \mu_j \in \mathbb{R}, \quad \sigma_j^2 \in (0, \infty),\ j = 1 \ldots K,\ \pi \in \triangle^{K-1}\}.$$

The variances of the components are assumed ordered, that is $\sigma_1^2 \leq \sigma_2^2 \leq \ldots \leq \sigma_K^2$, $\theta_k := (\mu_k, \sigma_k^2)$ denotes the coordinate projections on the state-dependent parameters for $1 \leq k \leq K$.

The compactification of both sets is done by adding limits of Cauchy sequences with respect to d_c as in Kiefer and Wolfowitz [27], and is denoted by $\bar{\Theta}^{full}$ and $\bar{\Theta}^{mix}$. Let $\alpha = (\alpha_1, \ldots, \alpha_K)$ be an initial state distribution, $\alpha_{i,j}$ the entries of Φ and $\varphi(y; \mu, \sigma^2)$ the density of the normal distribution with mean μ and variance σ^2:

$$\varphi(y; \mu, \sigma^2) = (2\pi)^{-\frac{1}{2}} \sigma^{-1} \exp(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}).$$

For $\theta \in \Theta^{full}$ the function

$$l_n^{full}(\theta; Y_1, \ldots, Y_n) = \log \sum_{x_1=1}^{K} \ldots \sum_{x_n=1}^{K} \alpha_{x_1} \varphi(Y_1, \theta_{x_1}) \prod_{t=2}^{n} \alpha_{x_{t-1},\, x_t} \varphi(Y_t, \theta_{x_t}) \tag{2.4}$$

is called the *log-likelihood function* for $Y_1, \ldots, Y_n$. For $\theta \in \Theta^{mix}$ the function

$$l_n^{mix}(\theta; Y_1, \ldots, Y_n) = \log \prod_{t=1}^{n} \sum_{j=1}^{K} \pi_j \varphi(Y_t, \theta_j) = \log \prod_{t=1}^{n} f(Y_t, \theta), \tag{2.5}$$

where $f(y; \theta) = \sum_{j=1}^{K} \pi_j \varphi(y; \theta_j)$, is called the *marginal-mixture-log-likelihood function* for $Y_1, \ldots, Y_n$.

Now penalty functions for the first stage of the procedure are defined similar to Chen et al. [12] .

Definition 2.1.2 A function $p_n : \Theta^{mix} \to \mathbb{R}$ with following properties:

1. $p_n(\theta) = \sum_{k=1}^{K} \tilde{p}_n(\sigma_k^2)$,
2. at any fixed θ, with $\sigma_k^2 > 0$, $k = 1, \ldots, K$, we have $p_n(\theta) = o(n)$, and $\sup_\theta \max\{0, p_n(\theta)\} = o(n)$,
3. p_n is differentiable and as $n \to \infty$, $p_n'(\theta) = o(n^{\frac{1}{2}})$ at any fixed σ_k^2, with $\sigma_k^2 > 0$, $k = 1, \ldots, K$,
4. for large enough n, $\tilde{p}_n(\sigma^2) \leq \sqrt{n} (\log n)^2 \log \sigma^2$, when $\sigma^2 < cn^{-2}$ for some $c > 0$,
5. for every $\varepsilon > 0$ holds $\sup_{\{\theta \mid \sigma^2(\theta) > \varepsilon\}} |\tilde{p}_n(\theta)| = o(n)$.

is called a *penalty function.*

These requirements are very similar to those from Chen et al. [12] and Chen and Tan [11]. The last condition was missing in the cited works, although it was implicitly assumed. The main difference lies in the fourth condition, which is linked to Lemma 2.2.9 below and is imposed to control the damaging effect of observations near degenerate components. Lemma 2.2.9 generalizes Lemma 1 from Chen and Tan [11] and is the most challenging part of the proof. The original proof relies on a Bernstein inequality for i.i.d. observations from Serfling [43], which is however not applicable for dependent observations. A more recent result from Merlevède et al. [35] was used instead.

The requirements are not very restrictive, for example the following function $\tilde{p}_n(\sigma^2) = -n^{-1}(\sigma^{-2} + \log \sigma^2)$ fulfils them.

Now we are ready to define the two-stage procedure for consistent parameter estimation.

Definition 2.1.3 Let

$$\hat{\theta}_n^{pIMLE} = \underset{\theta \in \Theta^{mix}}{\operatorname{argmax}}\, l_n^{mix}(\theta; Y_1, \ldots, Y_n) + p_n(\theta) \tag{2.6}$$

For ease of notation let $\nu(\theta) = (\mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)(\theta)$ for $\theta \in \Theta^{mix} \cup \Theta^{full}$ be the coordinate projection on the state-dependent parameters. For a mixture parameter $\theta' \in \Theta^{mix}$ and a $\delta > 0$ let

$$\Theta^{full}(\theta', \delta) = \{\theta \in \Theta^{full} \mid ||\nu(\theta), \nu(\theta')||_2 \leq \delta\}.$$

The *penalized maximum likelihood estimator (pMLE)* of θ is defined by

$$\hat{\theta}_n^{pMLE} = \underset{\theta \in \Theta^{full}(\hat{\theta}_n^{pIMLE},\, \delta)}{\operatorname{argmax}} l_n^{full}(\theta; Y_1, \ldots, Y_n) \tag{2.7}$$

for a penalty function p_n.

Now we are ready to establish the main result of this section, namely the consistency of the penalized maximum likelihood estimator for Gaussian hidden Markov models. The consistency is formulated in terms of the convergence in quotient topology (see Leroux [30]]).

Definition 2.1.4 For a parameter $\theta \in \Theta^{full}$, the *equivalence class* $\tilde{\theta}$ is defined by

$$\tilde{\theta} = \{\theta' \in \Theta^{full} \mid (\theta'_{X_i})_{i \in \mathbb{Z}} \overset{d}{=} (\theta_{X_i})_{i \in \mathbb{Z}}\},$$

that is the set of the parameters which induce the same law for the process $(\theta_{X_i})_{i\in\mathbb{Z}}$ as θ.

Convergence in quotient topology means that every open subset of the parameter space, that contains the equivalence class of θ_0, must for large n, contain the equivalence class of $\hat{\theta}_n^{pMLE}$.

Theorem 2.1.5 *$\hat{\theta}_n^{pMLE}$ a.s. converges to θ_0 in quotient topology with probability one for every positive $\delta > 0$ in the definition of $\hat{\theta}_n^{pMLE}$, for which $\Theta^{full}(\hat{\theta}_n^{pMLE}, \delta)$ does not contain any boundary point of Θ^{full}.*

The next theorem states asymptotic equivalence between the penalized MLE and the maximizer of the full HMM likelihood over a restricted parameter space, where the variances are bounded away from the zero. This allows us to transfer some results from the restricted case to the penalized one.

Theorem 2.1.6 (Asymptotic equivalence) *Denote the constrained maximizer $\hat{\theta}_R = \operatorname{argmax}_{\Theta^{full}} l_n^{full}(\theta; Y_1^n)$, s.t. $\sigma_k^2 \geq \varepsilon$, for $k \in \{1, \ldots, K\}$ for some small ε, such that ${\sigma_{0,k}}^2 > \varepsilon$, for $k \in \{1, \ldots, K\}$, then*

$$\sqrt{n}(\hat{\theta}_n^{pMLE} - \hat{\theta}_R) \xrightarrow{P} 0. \tag{2.8}$$

Proof. We expand $\nabla l_n^{full}(\hat{\theta}_n^{pMLE}) = \nabla l_n^{full}(\hat{\theta}_R) + \nabla^2 l_n^{full}(\tilde{\theta})(\hat{\theta}_n^{pMLE} - \hat{\theta}_R)$, where $\tilde{\theta}$ lies on the line segment between $\hat{\theta}_R$ and $\hat{\theta}_n^{pMLE}$. Since the true parameter lies in the interior of the feasible set, we have $\nabla l_n^{full}(\hat{\theta}_R) = 0$. So we obtain $\nabla l_n^{full}(\hat{\theta}_n^{pMLE}) = \nabla^2 l_n^{full}(\tilde{\theta})(\hat{\theta}_n^{pMLE} - \hat{\theta}_R)$. Furthermore, since $\hat{\theta}_n^{pMLE}$ and $\hat{\theta}_R$ are both consistent [1], we have $\tilde{\theta} \to \theta_0$ a.s.. Hence by the consistency of $\tilde{\theta}$ and Lemma 2 from Bickel et al. [8] it holds: $\frac{1}{n}\nabla^2 l_n^{mix}(\tilde{\theta}) \xrightarrow{P} -I_0$, where I_0 is a non-random matrix (the Fisher-Information) and by the continuous mapping theorem $n\nabla^2 l_n^{full}(\tilde{\theta})^{-1} \xrightarrow{P} -{I_0}^{-1}$. Combining these facts yields

$$\sqrt{n}(\hat{\theta}_n^{pMLE} - \hat{\theta}_R) = \overbrace{n\nabla^2 l_n^{full}(\tilde{\theta})^{-1}}^{\to -{I_0}^{-1}} \frac{1}{\sqrt{n}} \nabla l_n^{full}(\hat{\theta}_n^{pMLE}).$$

Finally it holds $\frac{1}{\sqrt{n}}\nabla l_n^{full}(\hat{\theta}_n^{pMLE}) \xrightarrow{P} 0$, since $\nabla l_n^{full}(\hat{\theta}_n^{pMLE}) = -\nabla p_n(\hat{\theta}_n^{pMLE})$ and $\frac{\partial}{\partial \theta_i} p_n(\hat{\theta}_n^{pMLE}) = o(\sqrt{n})$ a.s. by construction. □

[1] $\hat{\theta}_R$ satisfies conditions stated by Leroux [30]

The following result establishes the asymptotic normality of the penalized MLE.

Theorem 2.1.7 (Asymptotic normality)

$$\sqrt{n}(\hat{\theta}_n^{pMLE} - \theta_0) \xrightarrow{d} N(0, I_0^{-1}), \tag{2.9}$$

where $-I_0 = \lim_{n\to\infty} \frac{1}{n}\nabla^2 l_n^{full}(\theta_0, Y_1, \ldots, Y_n)$.

Proof. This statement follows from the asymptotic equivalence between $\hat{\theta}_n^{pMLE}$ and $\hat{\theta}_R$ and the fact, that $\hat{\theta}_R$ satisfies the assumptions of Theorem 1 in Bickel et al. [8]. The assumptions are:

(A1) The transition probability matrix is ergodic.

(A2) The elements of Φ and the stationary distribution are twice differentiable w.r.t θ.

(A3) Let $\theta = (\theta_1, \ldots, \theta_r)$. There exists $\delta > 0$, such that (i) for all $1 \leq i \leq r$ and all $k \in \{1, \ldots, K\}$

$$\mathbb{E}_0\left[\sup_{|\theta-\theta_0|<\delta} |\frac{\partial}{\partial\theta_i} \log \varphi(Y_1; \mu_k, \sigma_k^2)|^2\right] < \infty,$$

(ii) for all $1 \leq i, j \leq r$ and all $k \in \{1, \ldots, K\}$

$$\mathbb{E}_0\left[\sup_{|\theta-\theta_0|<\delta} |\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log \varphi(Y_1; \mu_k, \sigma_k^2)|\right] < \infty,$$

(iii) for all $j = 1, 2$, all $1 \leq i_l \leq r$, $l = 1, \ldots, j$, and all $k \in \{1, \ldots, K\}$

$$\int \sup_{|\theta-\theta_0|<\delta} |\frac{\partial^j}{\partial\theta_{i_1} \ldots \partial\theta_{i_j}} \varphi(Y_1; \mu_k, \sigma_k^2)| dy < \infty,$$

(A4) There exists $\delta > 0$ such that with

$$\rho_0(y) = \sup_{|\theta-\theta_0|<\delta} \max_{1\leq k_1,k_2\leq K} \frac{\varphi(y|\mu_{k_1}, \sigma_{k_1}^2)}{\varphi(y|\mu_{k_2}, \sigma_{k_2}^2)},$$

$\mathbb{P}(\rho_0(Y_1) = \infty \mid X_1 = k) < 1$ for all $k \in \{1, \ldots, K\}$.

(A5) θ_0 is an interior point of Θ

(A6) The maximum likelihood estimator is strongly consistent.

(A1) is part of our assumptions. The elements of Φ are part of the parameter vector and the initial distribution doesn't depend on θ, so (A2) is satisfied too. The conditions (A3) and (A4) are satisfied since φ is the normal density and $\sigma_k^2 > 0$ for $k \in \{1, \ldots, K\}$. Furthermore (A5) follows also from $\sigma_k^2 > 0$ for $k \in \{1, \ldots, K\}$. Finally (A6) holds, since $\hat{\theta}_R$ satisfies the regularity conditions from Leroux [30]. □

2.2 Proofs and technical results

Before a rigorous proof of Theorem 2.1.5 can be given, some general results on hidden Markov models, such as ergodicity and mixing properties, will be presented. First we deduce the Bernstein-type inequality (2.19) from Theorem 1 from Merlevède et al. [35]. Let us start by formulating a simplified version of that result.

Definition 2.2.1 Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{M}_1, \mathcal{M}_2 \subset \mathcal{A}$ sub-sigma-fields, $\mathcal{Z} = (Z_t)_{t\in\mathbb{Z}}$ real valued random variables.

1. The *α-dependence coefficient* between $\mathcal{M}_1$ and $\mathcal{M}_2$ is defined by

$$\alpha(\mathcal{M}_1, \mathcal{M}_2) = \sup\{|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| : A \in \mathcal{M}_1, B \in \mathcal{M}_2\} \quad (2.10)$$

2. For the sequence $(Z_i)_{i\in\mathbb{Z}}$ the *α-mixing (or strong-mixing) coefficient* is a function $\mathbb{N} \to \mathbb{R}^+$ defined by

$$\alpha_{\mathcal{Z}}(g) = \sup_{k\in\mathbb{N}} \alpha(\sigma(Z_t,\ -\infty < t \leq k),\, \sigma(Z_t,\ k+g \leq t < \infty)) \quad (2.11)$$

The conditions that are needed for the Bernstein inequality are the following. There exist positive constants a, b, γ_1 and c, $\gamma_2 > 0$ such that

$$\alpha(g) \leq a e^{-cg^{\gamma_1}}, \quad (B1)$$

$$\sup_t \mathbb{P}(|Z_t| > z) \leq e^{1-(z/b)^{\gamma_2}}, \quad (B2)$$

From Merlevède et al. [35], we have the following result.

Lemma 2.2.2 *Let $(Z_t)_{t\in\mathbb{Z}}$ be a sequence of centered real valued random variables, which satisfy Assumptions (B1) and (B2). Set $S_j = \sum_{t=1}^{j} Z_t$. Then there exist constants V, γ, C_1, C_2, C_3 and C_4 depending only on the constants a, b, γ_1 and c, $\gamma_2 > 0$ involved in Assumptions (B1) and (B2), such that for all $x > 0$,*

$$\mathbb{P}(\sup_{j\le n}|S_j| \ge x) \le n\exp\left(-\frac{x^\gamma}{C_1}\right) + \exp\left(-\frac{x^2}{C_2(1+nV)}\right) + \exp\left(-\frac{x^2}{C_3 n}\exp\frac{x^{\gamma(1-\gamma)}}{C_4(\log x)^\gamma}\right).$$

In order to use this result in the later proof (display (2.19)), we need to show that given a univariate Gaussian HMM $\mathcal{Y} = (Y_i)_{i\in\mathbb{Z}}$, the conditions (B1) and (B2) hold true for

$$\tilde{Z}^\tau_{t,k} = \mathbf{1}_{\{Y_t\le\eta_k+\tau\}} - \mathbf{1}_{\{Y_t\le\eta_{k-1}\}} - \big(F(\eta_k+\tau) - F(\eta_{k-1})\big), \tag{2.12}$$

where $\eta_k = F^{-1}(\frac{k}{n})$ and the constants a, b, γ_1 and c, $\gamma_2 > 0$ do not depend on k, τ and n, for every $n \in \mathbb{N}$. Since

$$\big|\tilde{Z}^\tau_{t,k}\big| \le 2 + 2M, \qquad \forall\ \tau \in (0, e^{-1}],\ 1 \le k \le n,\ n \ge 1,$$

this is evidently possible for (B2) and the constants b and γ_2. For (B1), we first consider the HMM itself. For lack of easy reference, we prove the following well-known result.

Proposition 2.2.3 *Let $\mathcal{Y} = (Y_t)_{t\in\mathbb{Z}}$ be a hidden Markov process with an irreducible and aperiodic underlying Markov chain. Then $\alpha(g) = \mathcal{O}(\rho^g)$ for some $0 < \rho < 1$.*

Proof. Since the process is assumed to be stationary, it suffices to show that

$$\sup\{|\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)| : A \in \sigma(Y_t;\ t \le 0),\ B \in \sigma(Y_t; t \ge g)\} \le c\rho^g \tag{2.13}$$

for some $c > 0$, $0 < \rho < 1$. First we prove (2.13) for certain algebras and then show that the sets, which satisfy (2.13) form a monotone class. An application of the monotone class theorem (e.g. Theorem 8.9 in Billingsley [9]) then completes the proof. We consider the following algebras

$$\mathcal{F}_0 = \{(Y_{i_1}, \dots, Y_{i_m}) \in B \mid B \in \mathcal{B}^m, -\infty < i_1, \dots, i_m < 0, m \in \mathbb{N}\},$$
$$\mathcal{F}_1 = \{(Y_{j_1}, \dots, Y_{j_l}) \in B \mid B \in \mathcal{B}^l, g \le j_1, \dots, j_l < \infty, l \in \mathbb{N}\}.$$

It is easy to see, that $\mathcal{F}_0$ and $\mathcal{F}_1$ are really algebras and generate $\sigma^2(Y_t, -\infty < t \le 0)$ and $\sigma^2(Y_t, g \le t < \infty)$ respectively. Now we assume $A \in \mathcal{F}_0$ and $B \in \mathcal{F}_1$,

that is there exist Borel sets B_1 and B_2 so that, $A = \{(Y_{i_1}, \ldots, Y_{i_m}) \in B_1\}$ and $B = \{(Y_{j_1}, \ldots, Y_{j_l}) \in B_2\}$ for some integer-vectors $(i_1, \ldots, i_m)$ and $(j_1, \ldots, j_l)$.

For $y \in \mathbb{R}$ we define $\tilde{\mathbb{P}}(y) = \operatorname{diag}(\varphi(y; \mu_1, \sigma_1^2), \ldots, \varphi(y; \mu_K, \sigma_K^2))$. With $\mathbf{1}$ we denote a column-vector of dimension K with 1 at every entry. Now we have

$$\begin{aligned}
\mathbb{P}(A)\mathbb{P}(B) &= \int_{B_1} \delta\tilde{\mathbb{P}}(y_1) \prod_{p=2}^{m} \Phi^{i_p - i_{p-1}} \tilde{\mathbb{P}}(y_p)\mathbf{1}dy \int_{B_2} \delta\tilde{\mathbb{P}}(y_1) \prod_{p=2}^{l} \Phi^{i_p - i_{p-1}} \tilde{\mathbb{P}}(y_p)\mathbf{1}dy \\
&= \int_{B_1 \times B_2} \delta\tilde{\mathbb{P}}(y_1) \prod_{p=2}^{m} \Phi^{i_p - i_{p-1}} \tilde{\mathbb{P}}(y_p)\mathbf{1}\delta\tilde{\mathbb{P}}(y_1') \prod_{p=2}^{l} \Phi^{j_p - j_{p-1}} \tilde{\mathbb{P}}(y_p')\mathbf{1}dydy' \\
\mathbb{P}(A \cap B) &= \int_{B_1 \times B_2} \delta\tilde{\mathbb{P}}(y_1) \prod_{p=2}^{m} \Phi^{i_p - i_{p-1}} \tilde{\mathbb{P}}(y_p)\Phi^{j_1 - i_m}\tilde{\mathbb{P}}(y_1') \prod_{p=2}^{l} \Phi^{j_p - j_{p-1}} \tilde{\mathbb{P}}(y_p')\mathbf{1}dydy'
\end{aligned}$$

We have $j_1 - i_m \geq g$ and from Theorem 8.9 in Billingsley [9] we have $\Phi^g \to \mathbf{1}\delta$ with exponential rate, that is $|\Phi^g - \mathbf{1}\delta| \leq c^*\rho^g\mathbf{1}\mathbf{1}'$. For some $c^* > 0$ and $0 < \rho < 1$. So we obtain

$$\begin{aligned}
&|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| = |\int_{B_1 \times B_2} \delta\tilde{\mathbb{P}}(y_1) \prod_{p=2}^{m} \Phi^{i_p - i_{p-1}} \tilde{\mathbb{P}}(y_p)\mathbf{1}\delta\tilde{\mathbb{P}}(y_1') \prod_{p=2}^{l} \Phi^{j_p - j_{p-1}} \tilde{\mathbb{P}}(y_p')\mathbf{1}dydy' \\
&- \int_{B_1 \times B_2} \delta\tilde{\mathbb{P}}(y_1) \prod_{p=2}^{m} \Phi^{i_p - i_{p-1}} \tilde{\mathbb{P}}(y_p)\Phi^{j_1 - i_m}\tilde{\mathbb{P}}(y_1') \prod_{p=2}^{l} \Phi^{j_p - j_{p-1}} \tilde{\mathbb{P}}(y_p')\mathbf{1}dydy'| \\
&= |\int_{B_1 \times B_2} \delta\tilde{\mathbb{P}}(y_1) \prod_{p=2}^{m} \Phi^{i_p - i_{p-1}} \tilde{\mathbb{P}}(y_p) \left(\mathbf{1}\delta - \Phi^{j_1 - i_m}\right) \tilde{\mathbb{P}}(y_1') \prod_{p=2}^{l} \Phi^{j_p - j_{p-1}} \tilde{\mathbb{P}}(y_p')\mathbf{1}dydy'| \\
&\leq \underbrace{\int_{B_1} \delta\tilde{\mathbb{P}}(y_1) \prod_{p=2}^{m} \Phi^{i_p - i_{p-1}} \tilde{\mathbb{P}}(y_p)dy}_{\leq \mathbf{1}'} \, |\left(\mathbf{1}\delta - \Phi^{j_1 - i_m}\right)| \, \underbrace{\int_{B_2} \tilde{\mathbb{P}}(y_1') \prod_{p=2}^{l} \Phi^{j_p - j_{p-1}} \tilde{\mathbb{P}}(y_p')\mathbf{1}dy'}_{\leq \mathbf{1}} \\
&\leq c^*\rho^g K^2
\end{aligned}$$

for every A, B of the assumed form. Here we used the convention $\int f dy = (\int f_1 dy, \ldots, \int f_K dy)$ for the integral of a vector-valued function f. Now, we have that for a fixed $B \in \mathcal{F}_1$, the set M_B of sets A satisfying that inequality builds a monotone class. Indeed, let $A_1 \subset A_2 \subset \ldots \subset A$, where $A_j \in M_B$. The measure $\mathbb{P}$ is continuous from below, so $|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| = |\mathbb{P}(\bigcup_{j=1}^{\infty} A_j \cap B) - \mathbb{P}(\bigcup_{j=1}^{\infty} A_j)\mathbb{P}(B)| = |\lim_{j\to\infty} P(A_j \cap B) - \lim_{j\to\infty} P(A_j)\mathbb{P}(B)| = \lim_{j\to\infty} |P(A_j \cap B) - P(A_j)\mathbb{P}(B)| \leq c\rho^g$. The same argument works for $A_1 \supset A_2 \supset \ldots \supset A$, since the measure $\mathbb{P}$ is also continuous from above. So M_A is a monotone class. By the

monotone class Theorem (Billingsley, Theorem 3.4) we can extend the inequality on the set $\sigma(\mathcal{F}_0) \times \mathcal{F}_1$. Now we fix an $A \in \sigma(\mathcal{F}_0)$ and the same argumentation applied to the set M_A of sets B satisfying the inequality for this A yields that also M_A is a monotone class. So finally we establish the inequality on the set $\sigma(\mathcal{F}_0) \times \sigma(\mathcal{F}_1)$. □

Lemma 2.2.4 *Given a univariate stationary Gaussian HMM, the variables $(\tilde{Z}^{\tau}_{t,k})$ in (2.12) satisfy the conditions (B1) and (B2), where the constants can be chosen independently of k and τ. Therefore, the Bernstein inequality in Lemma 2.2.2 applies, and all constants involved can be chosen independently of k and τ.*

Proof. We already discussed Assumption (B2) above. For (B1), since

$$\sigma(\tilde{Z}^{\tau}_{t,k};\ t \leq 0) \subset \sigma(Y_t;\ t \leq 0), \qquad \sigma(\tilde{Z}^{\tau}_{t,k};\ t \geq g) \subset \sigma(Y_t;\ t \geq g)$$

for any k and τ, the α-mixing coefficients are evidently uniformly bounded by those of the HMM. □

Stationarity affects marginal distributions of a process, while the strong mixing property describes the dependence intensity between process parts as function of the time gap between them. In the next lemma we combine the both properties to conclude ergodicity - a property which allows us to apply a strong law of large numbers to the process.

Lemma 2.2.5 *Let $(Y_i)_{i\in\mathbb{Z}}$ be a stationary strong mixing process. Then it is also ergodic.*

Proof. Since $(Y_t)_{t\in\mathbb{Z}}$ is a strong mixing process, we have for every $n, g \in \mathbb{N},\ A \in \sigma^2(Y^n_{-\infty}),\ B \in \sigma^2(Y^{\infty}_{n+g}) : |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)| < c\rho^g$ for a positive constant c and $0 < \rho < 1$. Now, let C be an invariant set, that is there exists a Borel set $B \in \mathcal{B}^{\mathbb{Z}}$, such that $C = \{T^{-k}Y^{\infty}_{-\infty} \in B\}$ for every $k \in \mathbb{N}$, where $T^0 = id,\ T^{-1}Y^{\infty}_{-\infty}(\omega)_n = Y_{n+1}(\omega),\ T^{-k} = T^{-(k-1)} \circ T^{-1}$. So T^{-1} is the left shift and T the right shift. According to Kolmogorov extension theorem, there is a sequence (C_n) of sets $C_n = \{Y^n_{-n} \in B_n\}$, for some cylinder set $B_n \in \mathcal{B}^{2n}$, such that $\mathbb{P}(C \,\Delta\, C_n) < 2^{-n}$, where $C \,\Delta\, C_n = \{C \setminus C_n\} \cup \{C_n \setminus C\}$ is the symmetric difference.
Now, since C is invariant and $(Y_i)_{i\in\mathbb{Z}}$ is stationary, we have

$$\mathbb{P}(T^{-k}C \,\Delta\, C_n) = \mathbb{P}(C \,\Delta\, T^k C_n) < 2^{-n},$$

for all $k, n \in \mathbb{N}$. Furthermore $T^k C_n = \{Y_{-n-k}^{n-k} \in B_n\}$, and hence $T^k C_n \in \sigma^2(Y_{-n-k}^{n-k}) \subset \sigma^2(Y_{-\infty}^{n-k})$ and $C_n \in \sigma^2(Y_{-n}^{n}) \subset \sigma^2(Y_{-n}^{\infty})$. Let $k \geq 2n$, $g_{k,n} = k - 2n$, then using the strong mixing property we conclude

$$|\mathbb{P}(C_n \cap T^k C_n) - \mathbb{P}(C_n)\mathbb{P}(T^k C_n)| < c\rho^{g_{k,n}},$$

for some $c > 0$ and $0 < \rho < 1$. We summarize, for every $\varepsilon > 0$ there exist $n, k \in \mathbb{N}$, such that

1. $||\mathbb{P}(C \cap C) - \mathbb{P}(C)^2| - |\mathbb{P}(C_n \cap T^k C_n) - \mathbb{P}(C_n)\mathbb{P}(T^k C_n)|| < \frac{\varepsilon}{2}$,
2. $|\mathbb{P}(C_n \cap T^k C_n) - \mathbb{P}(C_n)\mathbb{P}(T^k C_n)| < \frac{\varepsilon}{2}$,

therefore $|\mathbb{P}(C) - \mathbb{P}(C)^2| < \varepsilon$. Since $\varepsilon > 0$ was arbitrary, we have $\mathbb{P}(C) \in \{0, 1\}$. □

Now, we deduce some technical properties of the normal density.

Proposition 2.2.6 *Let $(\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$ set*

$$\tilde{A} = \tilde{A}(\mu, \sigma^2) = \{y \in \mathbb{R} \mid \frac{(y-\mu)^2}{\sigma^2} \leq (\log \sigma^2)^2\}. \tag{2.14}$$

Then

$$\varphi(y; \mu, \sigma^2) \leq \begin{cases} \sigma^{-1} & y \in \tilde{A} \\ \exp -\frac{(y-\mu)^2}{4\sigma^2} & \textit{otherwise.} \end{cases} \tag{2.15}$$

Proof. First we note $\varphi(y; \mu, \sigma^2) \leq \sigma^{-1}$ for every $y \in \mathbb{R}$, so the first inequality is obvious. For $y \notin \tilde{A}$ we have that $\frac{(y-\mu)^2}{\sigma^2} > (\log \sigma^2)^2$. Therefore

$$\begin{aligned} \varphi(y; \mu, \sigma^2) &\leq \frac{1}{\sqrt{\sigma^2}} \exp\big(-\frac{(y-\mu)^2}{\sigma^2}/4\big) \exp\big(-\frac{(y-\mu)^2}{\sigma^2}/4\big) \\ &< \frac{1}{\sqrt{\sigma^2}} \exp\big(-\frac{1}{4}(\log \sigma^2)^2\big) \exp\big(-\frac{(y-\mu)^2}{\sigma^2}/4\big) \\ &= \exp\Big(-\frac{1}{2}\big(\log \sigma^2 + (\log \sigma^2)^2/2\big)\Big) \exp(-\frac{(y-\mu)^2}{\sigma^2}/4) \\ &\leq \exp\big(-\frac{(y-\mu)^2}{\sigma^2}/4\big). \end{aligned} \tag{2.16}$$

□

Proposition 2.2.7 *Let $\mu_1, \mu_2 \in \mathbb{R}$ and $\sigma_1^2, \sigma_2^2 \in (0, \infty)$ with $\sigma_1^2 \leq \sigma_2^2 \leq \varepsilon$, for some $0 < \varepsilon < e^{-1/4}$. Suppose that $y \in \mathbb{R}$ is such that*

$$\frac{(y-\mu_1)^2}{\sigma_1^2} > (\log \sigma_1^2)^2, \qquad \frac{(y-\mu_2)^2}{\sigma_2^2} \leq (\log \sigma_2^2)^2.$$

then

$$\varphi(y; \mu_1, \sigma_1^2) < \varphi(y; \mu_2, \sigma_2^2).$$

Proof. From the properties of y we have

$$\frac{1}{\sqrt{\sigma_1^2}} \exp\{-\frac{1}{2}\frac{(y-\mu_1)^2}{\sigma_1^2}\} < \frac{1}{\sqrt{\sigma_1^2}} \exp\{-\frac{1}{2}(\log \sigma_1^2)^2\},$$
$$\frac{1}{\sqrt{\sigma_2^2}} \exp\{-\frac{1}{2}\frac{(y-\mu_2)^2}{\sigma_2^2}\} \geq \frac{1}{\sqrt{\sigma_2^2}} \exp\{-\frac{1}{2}(\log \sigma_2^2)^2\}.$$

Thus, it sufficies to show that the function

$$f(z) = \frac{1}{z} \exp\{-\frac{1}{2}\log(z^2)^2\}, \qquad z > 0,$$

is increasing near zero. The first derivative is given by

$$f'(z) = -\frac{1}{z^2} \exp\{-\frac{1}{2}\big(\log(z^2)\big)^2\} \left[1 + 4\log(z)\right],$$

which is > 0 for $z < e^{-1/4}$. □

Lemma 2.2.8 *Let Y be a random variable in $\mathbb{R}$ with a bounded density w.r.t. the Lebesgue measure. Given $\delta > 0$ there is a τ_0, such that for any $\mu \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$ with $\sigma^2 < \tau_0$, we have*

$$\mathbb{P}\big(Y \in \tilde{A}(\mu, \sigma^2)\big) < \delta,$$

where $\tilde{A}(\mu, \sigma^2)$ is defined in (2.14).

Proof. The Lebesgue length of $\tilde{A}(\mu, \sigma^2)$ is given by $2\sigma\,|\log \sigma^2|$, which tends to zero as $\sigma^2 \to 0$. The statement follows since Y has a bounded Lebesgue density. □

Bounds on the number of points near degenerate components. The following statement is related to (1.11). It bounds the number of observations of a Gaussian HMM process which are located in neighbourhoods of degenarate components. These observations have a high contribution to the likelihood and will be ruled out by the penalty function.

The difference is, that now intervals $(y, y+\tau]$ instead of ellipses are considered. Now, y plays the role of μ and τ the role of $|\Sigma|$ in (1.11).

Although intervals are simpler in their structure than ellipses, we can not follow the proof scheme as in the i.i.d. case via uniform law of iterated logarithm, since it assumed the independence of the observations. Instead, we generalize the proof from Chen et al. [12] via a Bernstein inequality from Merlevède et al. [35] and Borel-Cantelli lemma.

Lemma 2.2.9 *Let $(Y_t)_{t\in\mathbb{Z}}$ be a stationary Gaussian hidden Markov process with K states and parameter vector $(\Phi, \mu_1, \dots, \mu_K, \sigma_1^2, \dots, \sigma_K^2)$. Let F_n be the empirical distribution function of $Y_1, \dots, Y_n$, and M denote an upper bound for the marginal mixture density. Then almost sure there exists $N \in \mathbb{N}$, such that*

$$\sup_y [F_n(y+\tau) - F_n(y)] \leq \frac{(\log n)^2}{\sqrt{n}} + 2M\tau + \frac{1}{n}$$

for all $n \geq N$ and $\tau \in [0, e^{-1}]$.

Proof of Lemma 2.2.9. For $\tau = 0$ the statement is trivial. Let $\tau \in (0, e^{-1}]$ and $1 \leq k, i \leq n$ we define $\eta_k = F^{-1}(\frac{k}{n})$. We have

$$\begin{aligned}
&\sup_y [F_n(y+\tau) - F_n(y)] \\
&\quad \leq \max_k [F_n(\eta_k + \tau)] - F_n(\eta_{k-1}) \\
&\quad \leq \max_k [\{F_n(\eta_k + \tau) - F_n(\eta_{k-1})\} - \{F(\eta_k + \tau) - F(\eta_{k-1})\}] \\
&\quad + \max_k \{F(\eta_k + \tau) - F(\eta_{k-1})\}.
\end{aligned} \tag{2.17}$$

To bound the second term in (2.17), by the Mean Value Theorem we obtain

$$\begin{aligned}
F(\eta_k + \tau) - F(\eta_{k-1}) &= F(\eta_k + \tau) - F(\eta_k) + n^{-1} \\
&\leq M\tau + n^{-1} =: \delta_n(\tau).
\end{aligned} \tag{2.18}$$

It remains to find an appropriate bound for

$$\Delta_{n,k}^{\tau} = |\{F_n(\eta_k + \tau) - F_n(\eta_{k-1})\} - \{F(\eta_k + \tau) - F(\eta_{k-1})\}|.$$

Write

$$\begin{aligned} n\Delta_{n,k}^{\tau} &= |\sum_{t=1}^{n} \mathbf{1}_{\{Y_t \le \eta_k + \tau\}} - \mathbf{1}_{\{Y_t \le \eta_{k-1}\}} - \{F(\eta_k + \tau) - F(\eta_{k-1})\}| \\ &= |\sum_{t=1}^{n} Z_{t,k}^{\tau} - \{F(\eta_k + \tau) - F(\eta_{k-1})\}|. \end{aligned}$$

where $Z_{t,k}^{\tau} = \mathbf{1}_{\{Y_t \le \eta_k + \tau\}} - \mathbf{1}_{\{Y_t \le \eta_{k-1}\}}$. From the Bernstein inequality in Lemmas 2.2.2 and 2.2.4 there exist positive constants $\gamma, C_1, C_2, C_3, C_4, V$ and $n_0 \in \mathbb{N}$ depending only on the true parameter vector $(\Phi_0, \mu_{0,1}, \ldots, \mu_{0,k}, \sigma_{0,1}^2, \ldots, \sigma_{0,k}^2)$ of the HMM such that

$$\mathbb{P}(|\Delta_{n,j}^{\tau}| \ge x) \le n \exp\left(-\frac{n^{\gamma} x^{\gamma}}{C_1}\right) + \exp\left(-\frac{n^2 x^2}{C_2(1+nV)}\right) + \exp\left(-\frac{n^2 x^2}{C_3 n} \exp \frac{(nx)^{\gamma(1-\gamma)}}{C_4(\log\{xn\})^{\gamma}}\right) \tag{2.19}$$

for every $x \in \mathbb{R}$, $j = 1, \ldots, n$ and $\tau \in (0, e^{-1}]$. Setting $x = \frac{(\log n)^2}{2\sqrt{n}}$ gives

$$\begin{aligned} \mathbb{P}(|\Delta_{n,k}^{\tau}| \ge \frac{(\log n)^2}{2\sqrt{n}}) \le\ & n \exp\left(-\frac{n^{\frac{\gamma}{2}} (\log n)^{2\gamma}}{2^{\gamma} C_1}\right) \\ &+ \exp\left(-\frac{n(\log n)^4}{4C_2(1+nV)}\right) \\ &+ \exp\left(-\frac{(\log n)^4}{4C_3} \exp \frac{\{n^{\frac{1}{2}} (\log n)^2/2\}^{\gamma(1-\gamma)}}{C_4(\log\{(\log n)^2 n^{\frac{1}{2}}/2\})^{\gamma}}\right). \end{aligned}$$

Therefore we get that for every $n \ge n_0$, $j = 1, \ldots, n$ and $\tau \in (0, e^{-1}]$,

$$\mathbb{P}\Big(|\Delta_{n,k}^{\tau}| \ge \frac{(\log n)^2}{2\sqrt{n}}\Big) \le c n^{-3} \tag{2.20}$$

for some constant c. Let $r_n = \frac{(\log n)^2}{2M\sqrt{n}}$. It holds that

$$\begin{aligned} &\mathbb{P}\big(\max_{k=1\ldots n} |\Delta_{n,k}^{r_n}| \ge \frac{(\log n)^2}{2\sqrt{n}}\big) \le \mathbb{P}\Big(\cup_{k=1}^{n} \{|\Delta_{n,k}^{r_n}| \ge \frac{(\log n)^2}{2\sqrt{n}}\}\Big) \\ \le &\sum_{k=1}^{n} \mathbb{P}(|\Delta_{n,k}^{r_n}| \ge \frac{(\log n)^2}{2\sqrt{n}}) < c n^{-2}. \end{aligned} \tag{2.21}$$

By Borel-Cantelli, a.s. there is an N_1, such that

$$\max_{k=1\dots n} |\Delta_{n,k}^{r_n}| \leq \frac{(\log n)^2}{2\sqrt{n}}, \qquad n \geq N_1.$$

Therefore, by (2.17) and (2.18) and monotonicity,

$$\begin{aligned}\sup_{\tau\in(0,r_n]} \sup_y |F_n(y+\tau) - F_n(y)| &\leq \sup_y |F_n(y+r_n) - F_n(y)| \\ \leq \frac{(\log n)^2}{2\sqrt{n}} + \delta_n(r_n) &\leq \frac{(\log n)^2}{\sqrt{n}} + 1/n, \qquad n \geq N_1,\end{aligned}$$

which shows the estimate for all $\tau \in (0, r_n]$.
Next consider $\tau \in [r_n, e^{-1}]$. Now we define a finite grid over $[r_n, e^{-1}]$ by $\tau_0 = r_n$ and $\tau_{k+1} = 2\tau_k$, where $k \leq \lfloor \log_2 \frac{2Me^{-1}\sqrt{n}}{(\log n)^2} \rfloor =: k_n < \log n$ for n large enough. If $\tau_{k_n} < e^{-1}$, we add the point $\tau_{k_n+1} = e^{-1}$ to the grid, hence we assume w.l.o.g. $\tau_{k_n} = e^{-1}$. Let

$$D_n = \bigcup_{k=1}^{k_n} \Big\{ \sup_y F_n(y+\tau_k) - F_n(y) \geq \frac{(\log n)^2}{2\sqrt{n}} + \delta_n(\tau_k) \Big\}.$$

From (2.17), (2.18) and (2.20) we obtain

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{P}(D_n) &\leq \sum_{n=1}^{\infty} \sum_{j=1}^{k_n} \mathbb{P}\Big(\big\{ \sup_y F_n(y+\tau_j) - F_n(y) \geq \frac{(\log n)^2}{2\sqrt{n}} + M\tau_j + \frac{1}{n} \big\} \Big) \\ &\leq \sum_{n=1}^{\infty} \sum_{k=1}^{k_n} \mathbb{P}\Big(\max_{j=1\dots n} |\Delta_{n,j}^{\tau_k}| \geq \frac{(\log n)^2}{2\sqrt{n}} \Big) \\ &\leq \sum_{n=1}^{\infty} c \log n \, n^{-2} < \infty.\end{aligned}$$

where we estimate the maximal probability as in (2.21). We conclude by Borel-Cantelli $\mathbb{P}(D_n \text{ i.o.}) = 0$. Since for every $\tau \in [r_n, e^{-1}]$ there exist two grid points such that $\tau \in [\tau_j, \tau_{j+1}]$, a.s. there is an N_2 such that

$$\sup_y F_n(y+\tau) - F_n(y) \leq \sup_y F_n(y+\tau_{j+1}) - F_n(y) \leq \frac{(\log n)^2}{2\sqrt{n}} + 2M\tau + \frac{1}{n}$$

for all $n \geq N_2$ and $\tau \in [\tau_j, \tau_{j+1}]$, where we used $\tau_{j+1} \leq 2\tau$. □

Remark: The rate in the lemma above can be improved from $\sqrt{n}(\log n)^2$ to

$\sqrt{n}(\log n)^{1+q}$ For any $q > 0$. But the higher one is still sufficient for the proof.

Proof of Theorem 2.1.5 in case $K = 2$

Proof. It is sufficient to show the consistency of $\hat{\theta}_n^{pIMLE}$ for the state dependent parameters. Then the consistency of $\hat{\theta}_n^{pMLE}$ follows from the result in Leroux [30], since the maximization in stage 2 is carried out over a regular set, which contains the true parameter.

We show the consistency of $\hat{\theta}_n^{pIMLE}$ for the case $K = 2$ since the general K follows analogously. We follow Chen and Tan [11] in the proof structure and divide the parameter space into a finite number of subsets, one of which is regular. Step by step we show by applying Lemma 2.2.9 and classical techniques $\hat{\theta}_n^{pIMLE}$ to lie outside any of the irregular subsets.

In the following, the parameters μ_i, σ_i^2 will depend on θ, $i = 1, 2$, which we suppress in the notation.

Let $K = 2$ and assume w.l.o.g. $\sigma_1^2 \leq \sigma_2^2$. We divide the parameter space Θ^{mix} into three disjoint subsets.

$$\begin{aligned}
\Gamma_1 &= \{\, \theta \in \Theta^{mix} \,|\, \sigma_1^2 \leq \sigma_2^2 \leq \varepsilon_0 \,\}, \\
\Gamma_2 &= \{\, \theta \in \Theta^{mix} \,|\, \sigma_1^2 \leq \tau_0, \sigma_2^2 \geq \varepsilon_0 \,\}, \\
\Gamma_3 &= \Theta^{mix} \setminus \Gamma_1 \cup \Gamma_2.
\end{aligned}$$

For each $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \in \mathbb{R} \times \mathbb{R} \times (0, \infty) \times (0, \infty)$ we define the intervals subsets as in (2.14),

$$\tilde{A}_1 = \tilde{A}(\mu_1, \sigma_1^2), \qquad \tilde{A}_2 = \tilde{A}(\mu_2, \sigma_2^2).$$

Set

$$A_1 = \{t \mid Y_t \in \tilde{A}_1\}, \qquad A_2 = \{t \mid Y_t \in \tilde{A}_2\}, \tag{2.22}$$

and $M = \sigma_1^{-1}$. Further set

$$H_0 = \lim_n \frac{1}{n} l_n^{mix}(\theta_0^{mix}; Y_1^n), \tag{2.23}$$

which exists and is finite, see Lindgren [31]. The scalars ε_0 and τ_0 are chosen to satisfy

1. $2\sqrt{2}\varepsilon_0^{\frac{1}{2}} |\log \varepsilon_0| < e^{-1}, \qquad |\varepsilon_0^{\frac{1}{2}} \log \varepsilon_0 \log \varepsilon_0^{-\frac{1}{2}}| \leq 1/2.$

2. $0 < \tau_0 \leq \varepsilon_0$,
3. $-\log \varepsilon_0 - (\log \varepsilon_0)^2 \leq 4(H_0 - 2)$,
4. $\varepsilon_0 < \sigma_{01}^2$,
5. $\mathbb{P}(Y_1 \in \tilde{A}_1^c \cap \tilde{A}_2^c) \geq \frac{1}{2}$ for $\theta \in \Gamma_1$.

The first part of Condition 1 is necessary for applying Lemma 2.2.9, the second part is possible since $\varepsilon^{\frac{1}{2}} \log \varepsilon \log \varepsilon^{-\frac{1}{2}} \to 0$ as $\varepsilon \to 0$. The second condition ensures the order of the components. The third condition bounds the effect of observations, which will be ruled out by the log-likelihood at the true parameter. The existence of ε_0 and τ_0 which satisfy the first four conditions is obvious. The fifth condition can be achieved by applying Lemma 2.2.8.

Step 1. We shall show that

$$\sup_{\theta \in \Gamma_1} \left(l_n^{mix}(\theta; Y_1^n) + p_n(\theta) \right) - l_n^{mix}(\theta_0^{mix}; Y_1^n) - p_n(\theta_0^{mix}) \to -\infty. \tag{2.24}$$

To this end, we shall show that a.s. there is an $N \in \mathbb{N}$, such that for $n \geq N$ it holds that

$$l_n^{mix}(\theta; Y_1^n) + p_n(\theta) \leq n(H_0 - 1), \qquad n \geq N. \tag{2.25}$$

The conclusion then follows together with (2.23). To show (2.25), for a set $S \subset \{1, \ldots, n\}$ with $n(S)$ elements let

$$l_n^{mix}(\theta; S) = \log \prod_{j \in S} f(Y_j, \theta),$$

and write $l_n^{mix}(\theta; Y_1^n) + p_n(\theta) = \left(l_n^{mix}(\theta; A_1) + \tilde{p}_n(\sigma_1^2) \right) + \left(l_n^{mix}(\theta; A_1^c \cap A_2) + \tilde{p}_n(\sigma_2^2) \right) + l_n^{mix}(\theta; A_1^c \cap A_2^c)$. We shall bound each term on the right seperately in order to achieve (2.25). Since $\sigma_1^2 \leq \sigma_2^2$ we have that $f(y; \theta) \leq \sigma_1^{-1}$ for any y, and hence that $l_n^{mix}(\theta; A_1) \leq n(A_1) \log {\sigma_1}^{-1}$. First we assert for $\varepsilon_0 \geq \sigma_1^2 > n^{-2}$ with the help of Lemma 2.2.9

$$\begin{aligned} l_n^{mix}(\theta; A_1) &\leq n(A_1) \log \sigma_1^2 \leq \left(\sqrt{n} (\log n)^2 - nM\sigma_1 \log \sigma_1^2 + 1 \right) \log {\sigma_1}^{-1} \\ &= \sqrt{n} (\log n)^2 \log {\sigma_1}^{-1} - nM\sigma_1 \log \sigma_1 + \log {\sigma_1}^{-1} =: h_n(\sigma_1^2) \end{aligned}$$

and

$$\sup_{\sigma_1^2 \in [n^{-2},\, \varepsilon_0]} h_n(\sigma_1^2) \leq \sqrt{n} (\log n)^2 \log n - nM\varepsilon_0^{1/2} \log \varepsilon_0 + \log n < n/4. \tag{2.26}$$

The right hand side of the last display is less than a fraction of n for n large and ε small enough. Now, suppose $\sigma_1^2 \leq n^{-2}$, then from Property 4 of the penalty $\tilde{p}_n$ and Lemma 2.2.9, a.s. for large enough n, we obtain the bound

$$\begin{aligned} & l_n^{mix}(\theta; A_1) + \tilde{p}_n(\sigma_1^2) \leq n(A_1) \log {\sigma_1}^{-1} + \sqrt{n}(\log n)^2 \log \sigma_1^2 \\ \leq & \left(\sqrt{n}(\log n)^2 - nM\sigma_1 \log \sigma_1^2 + 1\right) \log {\sigma_1}^{-1} + \sqrt{n}(\log n)^2 \log \sigma_1^2 \\ = & \sqrt{n}(\log n)^2 \log \sigma_1 + \log {\sigma_1}^{-1} - nM\sigma_1 \log \sigma_1^2 \log {\sigma_1}^{-1} \\ \leq & n/4, \end{aligned} \tag{2.27}$$

since $\sqrt{n}(\log n)^2 \log \sigma_1 + {\sigma_1}^{-1}$ is negative, $\sigma_1^2 \leq \varepsilon_0$ and ε_0 is chosen to satisfy the second part of Condition 1 above. Similarly, for $y \in A_1^c \cap A_2$, from Lemma 2.2.7 we have that $f(y;\theta) \leq \log {\sigma_2}^{-1}$, and hence that $l_n^{mix}(\theta; A_1^c \cap A_2) \leq n(A_2) \log {\sigma_2}^{-1}$, and similarly as in (2.27) we obtain a.s. for large enough n that

$$\begin{aligned} \sup_{\sigma_1^2 \in [n^{-2},\, \varepsilon_0]} l_n^{mix}(\theta; A_1^c \cap A_2) + \tilde{p}_n(\sigma_2^2) \leq n/4 \\ \sup_{\sigma_1^2 \in (0,\, n^{-2})} l_n^{mix}(\theta; A_1^c \cap A_2) \leq n/4 \end{aligned} \tag{2.28}$$

Further,

$$\begin{aligned} l_n^{mix}(\theta; A_1^c \cap A_2^c) \leq & \sum_{j \in A_1^c \cap A_2^c} \log \left[\exp(\log {\sigma_2}^{-1} - \frac{1}{2}(\log \sigma_2^2)^2) \right] \\ \leq & \sum_{j \in A_1^c \cap A_2^c} -\frac{1}{2} \log \varepsilon_0 - \frac{1}{2}(\log \varepsilon_0)^2 \\ \leq & n(H_0 - 2). \end{aligned} \tag{2.29}$$

Here, for the first inequality we recall that the function $\frac{1}{z} \exp\{-\frac{1}{2} \log(z^2)^2\}$ is monotone increasing near zero, as shown in the proof of Lemma 2.2.7. Let us argue for the last inequality in (2.29). In case $H_0 < 2$, we assumed that $-\log \varepsilon_0 - (\log \varepsilon_0)^2 \leq 4(H_0 - 2)$, so that in this case we obtain

$$\begin{aligned} & \sum_{j \in A_1^c \cap A_2^c} -\frac{1}{2} \log \varepsilon_0 - \frac{1}{2}(\log \varepsilon_0)^2 \\ \leq & n(A_1^c \cap A_2^c)\, 2\, (H_0 - 2) \leq n(A_1^c \cap A_2^c)\, (H_0 - 2). \end{aligned}$$

In case $H_0 \geq 2$ we use the trivial bound $-\log \varepsilon_0 - (\log \varepsilon_0)^2 \leq 2(H_0 - 2)$, and get

$$\sum_{j \in A_1^c \cap A_2^c} -\frac{1}{2} \log \varepsilon_0 - \frac{1}{2}(\log \varepsilon_0)^2 \leq n(A_1^c \cap A_2^c)\, (H_0 - 2)$$

as well. By Condition 5 and the ergodic theorem, we get $n(A_1^c \cap A_2^c)/n \geq 1/2$ a.s., which gives the last estimate in (2.29). Now (2.25) follows from (2.26), (2.27), (2.28) and (2.29).

Step 2. Next, we show that

$$\sup_{\theta \in \Gamma_2} \Big(l_n^{mix}(\theta; Y_1^n) + p_n(\theta) \Big) - l_n^{mix}(\theta_0^{mix}) - p_n(\theta_0^{mix}) \to -\infty \; a.s.. \tag{2.30}$$

Define the set of indices $A_1 = A(\mu_1, \sigma_1^2)$ as in (2.22). We recall following bounds from the proof of Lemma 2.2.6

$$\varphi(y; \mu_1, \sigma_1^2) \leq \begin{cases} {\sigma_1}^{-1} \exp(-\frac{(\mu_1 - y)^2}{4\sigma_1^2}) & y \in \tilde{A}_1 \\ \exp(-\frac{(\mu_1 - y)^2}{4\sigma_1^2}) & \text{otherwise} \end{cases}.$$

Following Chen and Tan [11] we define a sub-density

$$g(y, \theta) = \pi_1 \exp(-\frac{(\mu_1 - y)^2}{4\sigma_1^2}) + \pi_2 \varphi(y; \mu_2, \sigma_2^2).$$

the function g is bounded by $\varepsilon_0^{-\frac{1}{2}}$ on Γ_2. Following statements hold for every $\theta \in \Gamma_2$:

$$\begin{aligned} &\log f(Y_t, \theta) \leq \log g(Y_t, \theta) + \mathbf{1}_{\{t \in A_1\}} \log {\sigma_1}^{-1}, \\ &l_n^{mix}(\theta) \leq n(A) \log {\sigma_1}^{-1} + \sum_{t=1}^{n} \log g(Y_t, \theta), \\ &\mathbb{E}_{\theta_0^{mix}} \log g(Y, \theta)/f(Y, \theta_0^{mix}) \leq \log \mathbb{E}_{\theta_0^{mix}} g(Y, \theta)/f(Y, \theta_0^{mix}) < 0, \\ &\frac{1}{n} \sum_{t=1}^{n} \log \frac{g(Y_t, \theta)}{f(Y_t, \theta_0^{mix})} \to \mathbb{E}_{\theta_0^{mix}} \log \frac{g(Y, \theta)}{f(Y, \theta_0^{mix})} < 0. \end{aligned}$$

Now, by using $\mathbb{E} \sup_{\theta \in U_\epsilon(\theta')} \varphi(y; \theta) < \infty$ for a sufficiently small neighborhood $U_\epsilon(\theta')$ of a $\theta' \in \Gamma_2$ and considering the compactification of Γ_2 by taking limits with respect to d_c, we apply the classical technique, see Wald [48], to obtain the statement $\lim_{n\to\infty} \sup_{\theta \in \Gamma_2} \frac{1}{n} \sum_{t=1}^{n} \log \frac{g(Y_t, \theta)}{f(Y_t, \theta_0^{mix})} =: -\kappa(\tau_0) < 0$, where $\kappa(\tau_0)$ is a decreasing

function, since larger τ_0 makes Γ_2 larger. Hence for a small enough $\tau_0 \leq \varepsilon_0$

$$\begin{aligned}
&\sup_{\theta\in\Gamma_2} l_n^{mix}(\theta) + p_n(\theta) - l_n^{mix}(\theta_0^{mix}) - p_n(\theta_0^{mix}) \\
&\quad\leq \sup_{\theta\in\Gamma_2} n(A)\log {\sigma_1}^{-1} + p_n(\theta) + \sum_{t=1}^{n} \log g(Y_t,\theta) - l_n^{mix}(\theta_0^{mix}) - p_n(\theta_0^{mix}) \\
&\quad\leq \sup_{\theta\in\Gamma_2} \left(\sqrt{n}(\log n)^2 - nM\sigma_1 \log \sigma_1^2 + 1\right) \log {\sigma_1}^{-1} + p_n(\theta) \\
&\quad+ \sup_{\theta\in\Gamma_2} \sum_{t=1}^{n} \log \frac{g(Y_t,\theta)}{f(Y_t,\theta_0^{mix})} - p_n(\theta_0^{mix}) \\
&\quad\leq \kappa(\varepsilon_0)n/2 - n\kappa(\varepsilon_0) = -\kappa(\varepsilon_0)n/2 - p_n(\theta_0^{mix}) \to -\infty.
\end{aligned}$$

We conclude $\hat{\theta}_n^{pIMLE} \in \Gamma_3$ which is regular and contains the true parameter θ_0^{mix}, so $\hat{\theta}_n^{pIMLE}$ is consistent for parameters of the stationary mixture.

The feasible set $\Theta^{full}(\hat{\theta}_n^{pIMLE},\ \delta)$ in stage 2 of the calculation of $\hat{\theta}_n^{pMLE}$ contains a.s. the true parameter θ_0 and is regular, so the consistency result from Leroux [30] can be applied. It completes the proof of the theorem. □

2.3 Conclusion

The existence of a consistent, asymptotically normal estimator for Gaussian hidden Markov models was proved. Ideas from the articles Chen et al. [12] and Chen and Tan [11] were used and generalized.

The proof was restricted to the one-dimensional case. The multivariate case could be proved if an analogon of Lemma 2.2.9 for more than one dimension would exist. In the i.i.d. setting such an analogon exists as shown in the section on penalized estimation of Gaussian mixture models, see Colorally 1.3.4. In order to obtain such a statement in the HMM setting, Alexander's law of iterated logarithm has to be generalized for dependent observations. This is an open issue and can be considered in future.

3 Identification of nonparametric hidden Markov models

In the following chapter we consider the problem of identification of nonparametric HMMs. Identification of HMMs is, like in the case of mixture models, an important issue, since it is a prerequisite for all subsequent statistical inference.

The mathematical formulation of the question is: does the distribution of the observed layer of a HMM; $(Y_t)_{t\in\mathbb{N}}$ determine the state-dependent distribution functions $y \mapsto \mathbb{P}(Y_1 \leq y \mid X_1 = k)$, $k = 1, \ldots, K$, the matrix of transition probabilities Φ and maybe even the initial distribution $\mathbb{P}_{X_1}$?

In an early work Petrie [39] considered HMMs with state dependent distributions having a finite support and characterized sets of identifiable parameters. Leroux [30] used a result from Teicher [45] on mixtures of product measures to prove identification of parametric HMMs under the assumption, that mixtures in the family of the state dependent distributions are identifiable.

Recently Allman et al. [6] proved identification results for some latent structure models including discrete HMMs. As a main tool thereby, they used Kruskal's result for identification of factors of triple products of matrices. Gassiat et al. [19] used the results from Allman et al. [6] to prove identification of nonparameteric HMMs, under the assumption, that the state-dependent distributions are linearly independent and the t.p.m. has full rank and is irreducible and aperiodic.

In this chapter we consider general hidden Markov models $(X_t, Y_t)_{t\geq 1}$. As before, we denote the entries of the t.p.m. by $\Phi = (\alpha_{j,k})_{j,k=1,\ldots,K}$. The conditional distributions of Y_t given $X_t = k$, $k = 1, \ldots, K$, are called the state-dependent distributions. We assume that they are independent of t. Further, assume that the Y_t take values in a subset of Euclidean space $\mathcal{S} = \mathbb{R}^d$, and denote the distribution functions of the state-dependent distributions by F_k, $k = 1, \ldots, K$. For $y \in \mathcal{S}$, $Y_t \leq y$ is meant componentwise.

We prove that the parameters $\Phi, F_1, \ldots F_K$ are identified up to relabeling from the distribution of $(2K + 1)$ consecutive observations of the HMM. If also the initial state probabilities of the Markov chain $\mathbb{P}_{X_1}(k)$, $1 \leq k \leq K$ have to be

identified, it suffices to know the distribution of $(2K+1)(K^2-2K+2)$ consecutive observations.

The only assumptions we need for this result are that the matrix Φ is ergodic and of full rank and the state-dependent distributions are distinct. This point distinguishes our work from the mentioned result from Gassiat et al. [19], where linear independence of the state-dependent distribution functions is required.

As Gassiat et al. [19], we also use the methodology from Allman et al. [6] to lead the problem back to the problem of identification of factors of triple products of matrices and to apply the powerful result from Kruskal [28].

Once identification is proved, positivity of the Kulbak-Leibler distance between two distinct HMMs based on blockwise likelihood follows immediately. Recovering construction from Leroux [30] yield this also for full-model likelihood.

The presented results are taken from the paper Alexandrovich and Holzmann [5].

3.1 Nonparametric identification

3.1.1 The stationary case

The following assumptions will be often used in our proofs.

A1 *The transition probability matrix* $\Phi = (\alpha_{j,k})_{j,k=1,\dots,K}$ *of* (X_t) *is irreducible, aperiodic and has a full rank.*

A2 *The state-dependent distributions* F_k, $k = 1, \dots, K$ *are all distinct.*

A3 (X_t) *is stationary and hence has the stationary starting distribution* π, *the stationary distribution of* Φ.

Let us first consider identification in the stationary case.

Theorem 3.1.1 *Suppose that for a known number of states* K, Φ *has a full rank,* $F_1, \dots, F_K$ *satisfy Assumption A2 and* $(X_t)_t$ *satisfies Assumption A3. Then the parameters* Φ *and* $F_1, \dots, F_K$ *are identified from the joint distribution of* $\left(Y_1, \dots, Y_{2K+1}\right)'$ *up to label swapping.*

It has to be emphasized that this statement is not implied by Theorem 1 in Gassiat et al. [19]. Assumption A2 is weaker than the assumption of linear independence and requires more elaborate arguments in the proof.

As it stands, the theorem only states that for given K, the parameters Φ and $F_1, \ldots, F_K$ are identified within the class of parameters satisfying imposed assumptions. However, from the proofs and exploiting the full strength of Kruskal's theorem, we easily get the following stronger result.

Corollary 3.1.2 *For given K, let Φ, $F_1, \ldots, F_K$ as well as $\tilde{\Phi}$, $G_1, \ldots, G_K$ be two sets of parameters for a K-state HMM, such that the joint distribution of an HMM $\left(Y_1, \ldots, Y_{2K+1}\right)'$ under both sets of parameters is equal. Further, suppose that Φ is regular, $F_1, \ldots, F_K$ satisfy Assumption A2 and $(X_t)_t$ satisfies Assumption A3. Then both sets of parameters coincide up to label swapping.*

Note that the Assumptions A1 and A2 are solely placed on Φ, $F_1, \ldots, F_K$, nothing is required for $\tilde{\Phi}$, $G_1, \ldots, G_K$.

3.1.2 General starting distribution

Now, let us turn to the case of a general starting distribution. This case is important for proving the definiteness of Kullback-Leibler divergence, based on the full-model likelihood, since there we also need identifiability of the initial distribution of the Markov chain. We need the following assumption:

A4 *(X_t) has the starting distribution λ.*

Now, the general identifiability result can be stated.

Theorem 3.1.3 *Suppose that for a known number of states K, Assumptions A1, A2 and A4 are satisfied. Then the parameters λ, Φ and $F_1, \ldots, F_K$ are identified from the joint distribution of $\left(Y_1, \ldots, Y_T\right)'$ with $T = (2K+1)(K^2 - 2K + 2) + 1$, up to label swapping.*

Similar to Corollary 3.1.2, this may be strengthened to the following result. The proof will be omitted, since it follows the same scheme as the proof of Corollary 3.1.2.

Corollary 3.1.4 *For given K, let λ, Φ, $F_1, \ldots, F_K$ as well as $\tilde{\lambda}, \tilde{\Phi}$, $G_1, \ldots, G_K$ be two sets of parameters for a K-state HMM (λ and $\tilde{\lambda}$ denote the starting distributions), such that the joint distribution of an HMM $\big(Y_1, \ldots, Y_T\big)'$ with $T = (2K+1)(K^2-2K+2)+1$ under both sets of parameters is equal. Further, suppose that Φ and $F_1, \ldots, F_K$ satisfy Assumptions A1 and A2. Then both sets of parameters coincide up to label swapping.*

3.1.3 Identifying the number of states

Before, we assumed the number of states as given a-priori. In fact, the power of Kruskal's theorem lets us identify the number of states as well.

Theorem 3.1.5 *Let λ, Φ and $F_1, \ldots, F_K$ be a set of parameters for a K-state HMM, and $\bar{\lambda}, \bar{\Phi}$ and $\bar{F}_1, \ldots, \bar{F}_L$ be a set of parameters for an L state HMM, where $L \leq K$. Assume that Φ satisfies A1 and that $F_1, \ldots, F_K$ satisfy A2. If the joint distribution of $\big(Y_1, \ldots, Y_T\big)$, $T = (2K+1)(K^2-2K+2)+1$ is the same under the both sets of parameters, then $K = L$ and the sets of parameters are equal up to a label swapping.*

Remark: Under a more restrictive assumption that also $\bar{\Phi}$ and $\bar{F}_1, \ldots, \bar{F}_L$ satisfy A1 and A2, the requirement $L \leq K$ could be omitted. Hence the number of states K is identified within the class of HMMs with ergodic and aperiodic transition probability matrices and distinct state-dependent distributions.

3.1.4 Kullback-Leibler distance of a HMM

In this section we indicate how the identification results can be used for nonparametric ML estimation.

Let ν be a σ-finite measure on $\mathcal{S}$, and let $\mathcal{D}$ be a class of densities on $\mathcal{S}$ w.r.t. ν.

Suppose that (Y_t, X_t) is a K-state HMM with t.p.m. Φ_0 satisfying Assumptions A1 and A3 having stationary distribution π_0, and that the state-dependent distributions $F_{0,1}, \ldots, F_{0,k}$ are all distinct and have densities $f_{0,1}, \ldots, f_{0,K}$ from the class $\mathcal{D}$.

First, we consider a blockwise likelihood function. For parameters λ, Φ, $f_1, \ldots, f_K$, $T \in \mathbb{N}$ and $\boldsymbol{y} = (y_1, \ldots, y_T)' \in \mathcal{S}^T$ consider

$$g_T\big(\boldsymbol{y}; \lambda, \Phi, f_1, \ldots f_K\big) = \sum_{x_1=1}^{K} \cdots \sum_{x_T=1}^{K} \lambda_{x_1} f_{x_1}(y_1) \prod_{i=2}^{T} \alpha_{x_{i-1},\, x_i} f_{x_i}(y_i),$$

the joint density w.r.t. $\nu^{\otimes T}$ of T observations under these parameters. Now,set

$$l_{T,n}\big(\lambda, \Phi, f_1, \ldots, f_K\big) = \sum_{i=0}^{n-1} \log g_T\big(Y_{iT+1}^{(i+1)T}; \lambda, \Phi, f_1, \ldots f_K\big),$$

a blockwise likelihood with blocklength T, which uses nT observations. From the ergodic theorem, we have a.s. that

$$\begin{aligned}\frac{1}{n}\Big(l_{T,n}\big(\lambda, \Phi, &f_1, \ldots, f_K\big) - l_{T,n}\big(\pi_0, \Phi_0, f_{0,1}, \ldots, f_{0,K}\big)\Big) \\ &\overset{n\to\infty}{\to} -KL\big(g_T(\cdot; \pi_0, \Phi_0, f_{0,1}, \ldots, f_{0,K}), g_T(\cdot; \lambda, \Phi, f_1, \ldots, f_K)\big) \leq 0,\end{aligned}$$

where KL is the Kullback-Leibler distance between the two densities on $\mathcal{S}^T$. If $T = (2K+1)(K^2 - 2K + 2) + 1$, Corollary 3.1.4 implies that this asymptotic contrast will identify the true parameter vector uniquely up to label swapping.

Now we show that the true parameter (except for the starting distribution) is also identified from the asymptotic contrast of the full-model log-likelihood, that is, the Kullback-Leibler distance of the HMM. We let

$$l_n\big(\lambda, \Phi, f_1, \ldots, f_K\big) = \log g_n\big(Y_1^n; \lambda, \Phi, f_1, \ldots f_K\big),$$

and impose in addition the following assumptions.

A5 $\mathbb{E}|\log f_{0,j}(Y_1)| < \infty$, $1 \leq j \leq K$

A6 $\mathbb{E}\,(\log f(Y_1))^+ < \infty$, *where* $f \in \mathcal{D}$.

Theorem 3.1.6 *Suppose that (Y_t, X_t) is a K-state HMM with t.p.m. Φ_0 satisfying Assumptions A1 and A3, and that the state-dependent distribution functions $F_{0,1}, \ldots, F_{0,k}$ are all distinct and have densities $f_{0,1}, \ldots, f_{0,K}$ from the class $\mathcal{D}$.*

Let Φ be a K-state t.p.m., let $f_1, \ldots, f_K \in \mathcal{D}$ and let λ, λ_0 be K-state probability vectors with strictly positive entries.

Furthermore let Assumptions A5 and A6 hold.

Then we have that a.s.,

$$\frac{1}{n}\Big(l_n\big(\lambda,\Phi,f_1,\dots,f_K\big)-l_n\big(\lambda_0,\Phi_0,f_{0,1},\dots,f_{0,K}\big)\Big) \\ \to -K\big((\Phi_0,f_{0,1},\dots,f_{0,K}),(\Phi,f_1,\dots,f_K)\big)\in(-\infty,0],$$

and $K\big((\Phi_0,f_{0,1},\dots,f_{0,K}),(\Phi,f_1,\dots,f_K)\big)=0$ *if and only if the two sets of parameters are equal up to label swapping.*

3.2 Proofs

3.2.1 Preliminaries

Let us recall a result of Kruskal in its precise form. For given matrices $M_i\in\mathbb{R}^{K\times n_i}$, $n_i\in\mathbb{N}$ $i=1,2,3$, let $A=\big\langle M_1,M_2,M_3\big\rangle$ denote the three-way array

$$A[i_1,i_2,i_3]=\sum_{k=1}^{K}(M_1)_{k,i_1}\,(M_2)_{k,i_2}\,(M_3)_{k,i_3},\qquad i_j=1,\dots,n_j,\quad j=1,2,3.$$

The Kruskal rank of a matrix $M\in\mathbb{R}^{K\times n}$, denoted $\operatorname{rank}_K M$, is the maximal j with $0\le j\le K$, for which each set of j rows in M are linearly independent (as vectors in $\mathbb{R}^n$).

Theorem A (Kruskal. Theorem 4a) *Let* $M_i,N_i\in\mathbb{R}^{K\times n_i}$, $n_i\in\mathbb{N}$ $i=1,2,3$ *be two sets of real matrices such that*

$$\big\langle M_1,M_2,M_3\big\rangle=\big\langle N_1,N_2,N_3\big\rangle.$$

Suppose that

$$rank_K\,M_1+rank_K\,M_2+rank_K\,M_3\ge 2K+2.$$

Then there exists a permutation matrix P *and diagonal matrices* $\Lambda_i\in\mathbb{R}^K$, *such that* $\Lambda_1\Lambda_2\Lambda_3=I$ *and*

$N_i=\Lambda_i\,PM_i,\quad i=1,2,3.$ ◇

Lemma 3.2.1 *If G_k, $k = 1, \dots, K$ are distinct distribution functions, then there exist a $t \in \mathbb{N}$ and $y_1 \dots, y_t \in \mathcal{S}$ such that the matrix $[(G_i(y_j))_{1 \le i \le K,\, 1 \le j \le t}, \mathbf{1}]$ has Kruskal rank at least two.*

Proof of Lemma 3.2.1. The distribution functions $G_1, \dots, G_K$ are distinct, hence for every pair $1 \le i < j \le K$ there exists $y \in \mathcal{S}$ such that $G_i(y) \neq G_j(y)$. Let $y_1, \dots, y_{\binom{K}{2}}$ be the points corresponding to $\binom{K}{2}$ pairs. Then the matrix

$$\left[(G_i(y_j))_{1 \le i \le K,\, 1 \le j \le \binom{K}{2}}, \mathbf{1}\right]$$

has Kruskal rank at least two. □

Now, we introduce a statement which leads to establishing the linear independence of the functions $P\big(Y_{T+2}^{2T+1} \le \cdot | X_{T+1} = k\big)$ (and related functions corresponding to the time reversal) under the imposed assumptions, that the transition matrix has full rank and the state-dependent functions are distinct. The proof goes by contradiction and uses some basic facts on dimensionality of orthogonal complements of certain subspaces.

Lemma 3.2.2 *Let $t \le K - 1$ and $v_1, \dots, v_t \in \mathbb{R}^K$ be linearly independent vectors. Assume that the entries of v_1 are all strictly positive. Let Φ be a $K \times K$ stochastic matrix of full rank and let $F_1, \dots, F_K$ be distinct distribution functions. Set*

$$D_y = \operatorname{diag}\big(F_1(y), \dots, F_K(y)\big).$$

Then there exists $y \in \mathcal{S}$ and a $1 \le j \le t$ for which the $K \times (t+1)$-matrix

$$\big[\Phi v_1, \dots, \Phi v_t, D_y \Phi v_j\big]$$

has full rank $t + 1$.

Proof of Lemma 3.2.2. First, we can construct vectors $o^{(1)}, \dots, o^{(K-t)} \in \mathbb{R}^K$ orthogonal to span $\{\Phi v_1, \dots, \Phi v_t\}$, which are of the form

$$o^{(i)} = (o_1^{(i)}, \dots, o_t^{(i)}, 0, \dots, -1, \dots, 0), \quad i = 1, \dots, K - t, \tag{3.1}$$

where the -1 is at the $t+i$'th place, after possibly relabeling the coordinates of $\mathbb{R}^K$. Indeed, observe that the $K \times t$ matrix $\Phi \cdot [v_1, \dots, v_t]$ has rank t, so that there are t linearly independent rows. Denote by M the $t \times t$ matrix formed from these rows,

and by N the $(K-t)\times t$ matrix consisting of the remaining rows, and assume (after relabeling) that

$$\Phi\cdot[v_1,\dots,v_t]=\begin{pmatrix}M\\N\end{pmatrix}.$$

For $e_i\in\mathbb{R}^{K-t}$ the i^{th} unit vector, we may set $o^{(i)}=\big(e_i'NM^{-1},-e_i'\big)'$.

Now, if there exist $y\in\mathcal{S}$, $1\le i\le K-t$ and $1\le j\le t$ for which $(D_y\Phi v_j)'o^{(i)}\neq 0$, then $D_y\Phi v_j$ cannot be contained in the t-dimensional subspace span$\{\Phi v_1,\dots,\Phi v_t\}$ of $\mathbb{R}^K$, and the assertion of the lemma follows.

Thus assume that

$$(D_y\Phi v_j)'o^{(i)}=0,\qquad \forall\ y\in\mathcal{S},\ 1\le i\le K-t,\ 1\le j\le t, \tag{3.2}$$

this will lead to a contradiction. Let $\gamma_1,\dots,\gamma_K$ denote the row vectors of Φ. For $i=1,\dots,K-t$ set

$$S_i:=\text{span}\big\{o_1^{(i)}F_1(y)\gamma_1+\ldots+o_t^{(i)}F_t(y)\gamma_n-F_{t+i}(y)\gamma_{t+i}\mid y\in\mathcal{S}\big\}.$$

Then (3.2) implies that

$$\text{span}\big\{S_1,\dots,S_{K-t}\big\}\subseteq\text{span}\big\{v_1,\dots,v_t\big\}^{\perp}. \tag{3.3}$$

We first argue that if (3.3) holds,

$$\dim S_i\ge 2,\quad i=1,\dots,K-t. \tag{3.4}$$

To this end we assert that among the first t elements of $o^{(i)}$ there is at least one non-zero entry. Indeed, suppose that all n entries were equal zero, then by the construction of $o^{(i)}$, definition of S_i and (3.3), we get that

$$F_{t+i}(y)\gamma_{t+i}'v_1=0\quad\forall\ y\in\mathcal{S},$$

a contradiction since $\gamma_{t+i}'v_1>0$ (since we assume that v_1 has strictly positive entries).

Thus, assume that $j\in\{1,\dots,t\}$ is such that $o_j^{(i)}\neq 0$. Since F_j and F_{t+i} are distinct distribution functions, there exist $y_1^{(i)},y_2^{(i)}$ such that the vectors

$$\big(F_j(y_1^{(i)}),F_{t+i}(y_1^{(i)})\big)\quad\text{and}\quad\big(F_j(y_2^{(i)}),F_{t+i}(y_2^{(i)})\big)$$

are linearly independent, and hence so are the vectors

$$\big(o_1^{(i)}F_1(y_l^{(i)}),\dots,o_t^{(i)}F_t(y_l^{(i)}),-F_{t+i}(y_l^{(i)})\big),\quad l=1,2,$$

of coefficients of the linearly independent vectors $\gamma_1,\dots,\gamma_t,\gamma_{t+i}$, which shows (3.4).

To conclude the proof, we observe that due to the linear independence of $\gamma_1,\dots,\gamma_K$ and the definition of the S_i, we have that

$$S_i \not\subseteq \operatorname{span}\Big\{\bigcup_{j=1,\,j\neq i}^{K-t} S_j\Big\}\qquad \forall\ i=1,\dots,K-t.$$

Together with (3.4) we obtain that

$$\dim\Big(\operatorname{span}\big\{S_1,\dots,S_{K-t}\big\}\Big)\geq K-t+1,$$

a contradiction to (3.3). This concludes the proof of the lemma. □

3.2.2 Proofs for Section 3.1.1

Proof of Theorem 3.1.1. Step 1: Linear independence. Let $T\geq K-1$, and consider

$$V_T=Y_1^T=\big(Y_1,\dots,Y_T\big)',\qquad W_T=Y_{T+2}^{2T+1}=\big(Y_{T+2},\dots,Y_{2T+1}\big)'.$$

The conditional distribution functions of W_T given $X_{T+1}=k$, $k=1,\dots,K$, are given by

$$\begin{aligned}P\big(W_T\leq \boldsymbol{y}|X_{T+1}=k\big)&=\sum_{k_1=1}^{K}\dots\sum_{k_T=1}^{K}\alpha_{kk_1}\prod_{t=2}^{T}\alpha_{k_{t-1}k_t}\prod_{t=1}^{T}F_{k_t}(y_t)\\&=:G_T(\boldsymbol{y};k),\qquad \boldsymbol{y}=(y_1,\dots,y_T)'\in\mathcal{S}^T.\end{aligned}$$

Lemma 3.2.3 *Under Assumptions A1 and A2, for $T\geq K-1$ the distribution functions $G_T(\cdot;k)$, $k=1,\dots,K$, are linearly independent.*

Proof of Lemma 3.2.3. Since marginal distributions of linearly dependent distributions remain linearly dependent, it is enough to show linear independence for $T=K-1$.

Now, we construct points $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K \in \mathcal{S}^{K-1}$, for which the $K \times K$ matrix $\big(G_{K-1}(\boldsymbol{y}_t; k)\big)_{k,t=1,\ldots,K}$ has full rank K (k is the row index and t the column index).

For $\boldsymbol{y} = (y_1, \ldots, y_t)' \in \mathcal{S}^t$ consider

$$\begin{aligned}\tilde{G}_t(\boldsymbol{y}; k) &= F_k(y_1) \sum_{k_2=1}^{K} \ldots \sum_{k_t=1}^{K} \alpha_{kk_2} \prod_{s=2}^{t-1} \alpha_{k_s k_{s+1}} \prod_{s=2}^{t} F_{k_s}(y_s) \\ &= F_k(y_1)\, \gamma_k D_{y_2}\, \Phi \ldots \Phi D_{y_t}\, (1, \ldots, 1)', \quad k = 1, \ldots, K,\end{aligned}$$

where as above, $D_y = \operatorname{diag}\big(F_1(y), \ldots, F_K(y)\big)$ and γ_k are the row vectors of Φ. Since

$$\big(\tilde{G}_{K-1}(\boldsymbol{y}_t; k)\big)_{k,t=1,\ldots,K} = \Phi \cdot \big(G_{K-1}(\boldsymbol{y}_t; k)\big)_{k,t=1,\ldots,K},$$

it is enough to find $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K \in \mathcal{S}^{K-1}$ for which $\big(\tilde{G}_{K-1}(\boldsymbol{y}_t; k)\big)_{k,t=1,\ldots,K}$ has full rank K. We show by induction:

Claim: For $t = 1, \ldots, K-1$ there exist vectors $\boldsymbol{y}_1^{(t)}, \ldots, \boldsymbol{y}_{t+1}^{(t)} \in \mathcal{S}^t$ for which the vectors

$$v_j^{(t)} = \begin{pmatrix} \tilde{G}_t(\boldsymbol{y}_j^{(t)}; 1) \\ \vdots \\ \tilde{G}_t(\boldsymbol{y}_j^{(t)}; K) \end{pmatrix}, \quad j = 1, \ldots, t+1,$$

are linearly independent, and $v_1^{(t)}$ has only strictly positive entries.
The case $t = K-1$ will establish the lemma.

Proof of Claim. For $t = 1$, we find $y_1^{(1)}, y_2^{(1)} \in \mathcal{S}$ for which

$$v_j^{(1)} = \big(F_1(y_j^{(1)}), \ldots, F_K(y_j^{(1)})\big), \quad j = 1, 2,$$

are linearly independent, and for which $v_1^{(1)}$ has only positive entries. Now, suppose that the claim is valid for t. We apply Lemma 3.2.2 and find a $y_0 \in \mathcal{S}$ and a $1 \le j \le t+1$ for which the $K \times (t+2)$ matrix

$$M := \big[\Phi v_1^{(t)}, \ldots, \Phi v_{t+1}^{(t)},\ D_{y_0}\, \Phi v_j^{(t)}\big]$$

has full rank $t+2$, which means that it has a $(t+2) \times (t+2)$ submatrix of non-zero determinant. Since $D_y \to I_K$, as all coordinates of y tend to ∞,

$$\big[D_y \Phi v_1^{(t)}, \ldots, D_y \Phi v_{t+1}^{(t)},\ D_{y_0}\, \Phi v_j^{(t)}\big] \to M,$$

and the corresponding submatrix will also be of non-zero determinant for an appropriate $y \in \mathcal{S}$ (for which also D_y has positive entries on its diagonal). The claim

for $t+1$ now follows by setting

$$\boldsymbol{y}_s^{(t+1)} := \big(y, (\boldsymbol{y}_s^{(t)})'\big)', \quad s = 1, \dots, t+1, \quad \text{and} \quad \boldsymbol{y}_{t+2}^{(t+1)} = \big(y_0, (\boldsymbol{y}_j^{(t)})'\big)'.$$

□

Similarly, consider the time reversal $\tilde{\Phi} = \big(\tilde{\alpha}_{j,k}\big)_{j,k=1,\dots,K}$ with

$$\tilde{\alpha}_{j,k} = \frac{\pi_k \alpha_{k,j}}{\pi_j}.$$

Then

$$\begin{aligned} P\big(V_T \leq (y_T, \dots, y_1)' | X_{T+1} = k\big) &= \sum_{k_1=1}^{K} \dots \sum_{k_T=1}^{K} \tilde{\alpha}_{kk_1} \prod_{t=1}^{T-1} \tilde{\alpha}_{k_t k_{t+1}} \prod_{t=1}^{T} F_{k_t}(y_t) \\ &=: H_T(y_T, \dots, y_1; k). \end{aligned}$$

Lemma 3.2.4 *Under Assumptions A1, A2 and A3, $T \geq K-1$ the distribution functions $H_T(\cdot; k)$, for $k = 1, \dots, K$, are linearly independent.*

This is immediate since $\tilde{\Phi}$ has full rank as well.

Step 2: Identification of conditional distributions.

Lemma 3.2.5 *Under Assumptions A1 - A3, for $T \geq K-1$ we identify the distribution functions $H_T(; k), F_k, G_T(; k)$, $k = 1, \dots, K$, up to joint label swapping.*

Proof of Lemma 3.2.5. From the proof of Lemma 3.2.1 we know that there exist points $y_j \in \mathcal{S}$, $j = 1, \dots, \binom{K}{2}$, such that the matrix

$$M_2 := \Big[\big(F_i(y_j)\big)_{1 \leq i \leq K,\, 1 \leq j \leq K(K-1)/2}, \mathbf{1} \Big], \tag{3.5}$$

where $\mathbf{1}$ is a K-dimensional column-vector consisting of ones, has Kruskal rank at least 2.

From Lemma 3.2.3, we may choose $\boldsymbol{y}_1, \dots, \boldsymbol{y}_K \in \mathcal{S}^T$ such that the $K \times (K+1)$-matrix

$$M_3 := \Big[\big(G_T(\boldsymbol{y}_t; k)\big)_{1 \leq k \leq K,\, 1 \leq t \leq K}, \mathbf{1} \Big] \tag{3.6}$$

has full rank K, see Lemma 17 in Allman et al. [6] or the argument in Step 1. Similarly accordingly to Lemma 3.2.4 we find $\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_K \in \mathcal{S}^T$ such that the $K \times (K+1)$-matrix

$$M_1 := \left[(H_T(\tilde{\boldsymbol{y}}_t; k))_{1 \le k \le K,\, 1 \le t \le K}, \mathbf{1} \right] \tag{3.7}$$

has rank K. Let $\tilde{M}_1 := \mathrm{diag}(\pi) M_1$. The matrix $\tilde{M}_1$ still has full rank, since $\pi_k > 0$ for $k = 1, \ldots, K$.

We conclude that

$$\mathrm{rank}_K(\tilde{M}_1) + \mathrm{rank}_K(M_2) + \mathrm{rank}_K(M_3) = 2K + 2, \tag{3.8}$$

where rank_K denotes the Kruskal rank of a matrix.

Now, consider the triple product $M := \langle \tilde{M}_1, M_2, M_3 \rangle$, which is defined by

$$M[i,j,r] = \sum_{k=1}^{K} (\tilde{M}_1)_{k,i} \, (M_2)_{k,j} \, (M_3)_{k,r},$$

where $1 \le i, r \le K+1$ and $1 \le j \le \binom{K}{2} + 1$. We show that M is identified from the joint distribution of Y_1^{2T+1}. First, consider $1 \le i, r \le K$ and $1 \le j \le \binom{K}{2}$

$$\begin{aligned}
&M[i,j,r] = \sum_{k=1}^{K} \pi_k \, H_T(\tilde{\boldsymbol{y}}_i; k) F_k(y_j) G_T(\boldsymbol{y}_r; k) \\
=&\sum_{k=1}^{K} \pi_k P(Y_1^T \le \tilde{\boldsymbol{y}}_i | X_{T+1} = k) P(Y_{T+1} \le y_j | X_{T+1} = k) P(Y_{T+2}^{2T+1} \le \boldsymbol{y}_r | X_{T+1} = k) \\
=&\sum_{k=1}^{K} \pi_k P(Y_1^T \le \tilde{\boldsymbol{y}}_i, Y_{T+1} \le y_j, Y_{T+2}^{2T+1} \le \boldsymbol{y}_r | X_{T+1} = k) \\
=&P(Y_1^T \le \tilde{\boldsymbol{y}}_i, Y_{T+1} \le y_j, Y_{T+2}^{2T+1} \le \boldsymbol{y}_r)
\end{aligned} \tag{3.9}$$

Similarly, setting $m = \binom{K}{2}$,

$$\begin{aligned}
&M[K+1,j,r] = P(Y_{T+1} \le y_j, Y_{T+2}^{2T+1} \le \boldsymbol{y}_r), && M[K+1,m+1,r] = P(Y_{T+2}^{2T+1} \le \boldsymbol{y}_r) \\
&M[i,m+1,r] = P(Y_1^T \le \tilde{\boldsymbol{y}}_i, Y_{T+2}^{2T+1} \le \boldsymbol{y}_r), && M[K+1,j,K+1] = P(Y_{T+1} \le y_j), \\
&M[i,j,K+1] = P(Y_1^T \le \tilde{\boldsymbol{y}}_i, Y_{T+1} \le y_j), && M[i,m+1,K+1] = P(Y_1^T \le \tilde{\boldsymbol{y}}_i),
\end{aligned} \tag{3.10}$$

as well as $M[K+1, m+1, K+1] = 1$. Evidently, all of these quantities are identified from the distribution of Y_1^{2T+1}.

Now, using (3.8) we apply Theorem A to show that the matrices $\tilde{M}_1, M_2$ and M_3 are identified from M up to scaling and permutation, that is there exist a permutation matrix P and diagonal matrices $\Lambda_1, \Lambda_2, \Lambda_3$, such that $\Lambda_1 P\tilde{M}_1, \Lambda_2 P M_2$, and $\Lambda_3 P M_3$ are known and the relationship $\Lambda_1\Lambda_2\Lambda_3 = I$ holds.

Since we know that in the last column of M_2 there are only ones, we obtain the i^{th} diagonal element of the scaling matrix Λ_2 as $(\Lambda_2 P M_2)_{i,K+1}$ for each $i = 1, \ldots, K$. Similarly we find the matrix Λ_3. The elements of Λ_1 can then be determined by the relationship $\Lambda_1\Lambda_2\Lambda_3 = I_K$. Hence we identified the matrices $\tilde{M}_1, M_2$ and M_3 up to simultaneous row permutations.

In order to identify the values of $H_T(\boldsymbol{y}; k), F_k(y), G_T(\tilde{\boldsymbol{y}}; k)$ at any arbitrary points $\boldsymbol{y}, \tilde{\boldsymbol{y}} \in \mathcal{S}^T, y \in \mathcal{S}$, we insert the corresponding columns into matrices $\tilde{M}_1, M_2$ and M_3 respectively without changing the validity of (3.8). □

Step 3: Identification of Φ.

We choose $T = K - 1$, and after applying the result in Step 2, fix a labeling $H_T(; k), F_k, G_T(; k)$, $k = 1, \ldots, K$. It remains to identify the t.p.m. Φ.

Again, we choose $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K \in \mathcal{S}^T$ such that the $K \times K$-matrix

$$A_1 = \big(G_T(\boldsymbol{y}_t; k)\big)_{k=1,\ldots,K; t=1,\ldots,K}$$

has full rank K, see Lemma 17 in Allman et al. (2009) or the argument in Step 1. For $y \in \mathcal{S}$ consider the $K \times K$-matrix

$$A_2 = \big(G_{T+1}((y, \boldsymbol{y}_t')'; k)\big)_{k=1,\ldots,K; t=1,\ldots,K}.$$

From Step 2, $H_{T+1}(; k), F_k, G_{T+1}(; k)$, $k = 1, \ldots, K$, and hence A_2 are identified up to joint label swapping. Since the F_k are all distinct, we may choose the same labeling as the one fixed for $H_T(; k), F_k, G_T(; k)$, $k = 1, \ldots, K$. In this case, we have that

$$A_2 = \Phi \operatorname{diag}\big(F_1(y), \ldots, F_K(y)\big) A_1.$$

Choose y so that $F_k(y) \neq 0$, $k = 1, \ldots, K$, so that Φ is identified as

$$\Phi = A_2\, A_1^{-1} \operatorname{diag}\big(F_1(y)^{-1}, \ldots, F_K(y)^{-1}\big).$$

□

Proof of Corollary 3.1.2. We choose $y_j \in \mathcal{S}$, $j = 1, \dots, \binom{K}{2}$, $\boldsymbol{y}_1, \dots, \boldsymbol{y}_K \in \mathcal{S}^T$, $\tilde{\boldsymbol{y}}_1, \dots, \tilde{\boldsymbol{y}}_K \in \mathcal{S}^T$, such that the matrices M_j, $j = 1, 3$ in (3.6) and (3.7) have full rank, and that the matrix M_2 in (3.5) has Kruskal rank 2 for the parameters Φ and $F_1, \dots, F_K$, and let $\tilde{M}_1 = \operatorname{diag}(\pi) M_1$, where π is the stationary distribution of Φ.

Define the matrices N_1, N_3 and N_2 in a similar way for the parameter sets $\tilde{\Phi}$ and $\tilde{F}_1, \dots, \tilde{F}_K$. If its starting distribution is δ, consider $\tilde{N}_1 = \operatorname{diag}(\delta\Phi^{K-1}) N_1$ ($\delta\Phi^{K-1}$ is the marginal distribution of X_K under this parameter set).

Now, (3.9) and (3.10) show that under the assumption that both sets of parameter sets induce the same distribution of $Y_1, \dots, Y_{2K-1}$,

$$\langle \tilde{M}_1, M_2, M_3 \rangle = \langle \tilde{N}_1, N_2, N_3 \rangle.$$

From Kruskal's Theorem A, there is a $K \times K$ permutation matrix P and diagonal matrices Λ_i, $i = 1, 2, 3$ with $\Lambda_1\Lambda_2\Lambda_3 = I_K$, such that

$$M_i = \Lambda_i P N_i, \quad i = 2, 3, \qquad \tilde{M}_1 = \Lambda_1 P \tilde{N}_1.$$

Since M_i, N_i $i = 2, 3$, have only ones in the last column, $\Lambda_2 = \Lambda_3 = I_K$ and hence also $\Lambda_1 = I_K$. It follows that N_3 and $\tilde{N}_1$ must also have full rank, and that P is uniquely determined.

If we insert columns with entries $H_T(\boldsymbol{y}; k), F_k(y), G_T(\tilde{\boldsymbol{y}}; k)$ at any arbitrary points $\boldsymbol{y}, \tilde{\boldsymbol{y}} \in \mathcal{S}^T, y \in \mathcal{S}$, the matrix P must be the same, so that we get the equality of F_k and G_k, up to label swapping. Then arguing as in Step 3 of Theorem 3.1.1, the matrices A_1 and A_2 must be equal for both sets of parameters up to permutation of rows, which shows that $\Phi = P\tilde{\Phi}$ for a permutation matrix P.

□

3.2.3 Proofs for Sections 3.1.2, 3.1.3 and 3.1.4

Proof of Theorem 3.1.3. Step 1.: First assume that λ has only positive entries. Then from the joint distribution of $\big(Y_1, \dots, Y_{2K+1}\big)$ we identify Φ and $F_1, \dots, F_K$, as well as the conditional distributions

$$H_T(\boldsymbol{y}; k) := P\big(Y_1^T \leq \boldsymbol{y} \,|\, X_{T+1} = k\big), \qquad k = 1, \dots, K, \ \boldsymbol{y} = (y_1, \dots, y_T)' \in \mathcal{S}^T.$$

for $T = K - 1, K$, up to label swapping.

Proof of claim of Step 1. We may follow the proof of Theorem 3.1.1, and it suffices to show that the distribution functions $H_T(\cdot;k)$, $k=1,\dots,K$, are linearly independent, where $T=K-1$. The time reversal

$$(X_{T+1},\dots,X_1)$$

is an inhomogeneous Markov chain, and therefore $\big((X_{T+1},Y_{T+1}),\dots,(X_1,Y_1)\big)$ is an HMM with inhomogeneous underlying Markov chain and state-dependent distributions $F_1,\dots,F_K$. More precisely, letting

$$\lambda^{(t)}=\lambda\Phi^{t-1},\qquad \big(\tilde\Phi^{(t)}\big)_{i,j}=\frac{\lambda_j^{(t)}\alpha_{j,i}}{\sum_{k=1}^K\lambda_k^{(t)}\alpha_{k,i}}=:\big(\tilde\alpha_{i,j}^{(t)}\big)_{i,j=1,\dots,K},\quad t=1,\dots,T,$$

we have that for $\boldsymbol{y}=(y_1,\dots,y_T)'\in\mathcal{S}^T$ that

$$H_T(\boldsymbol{y};k)=\sum_{k_1=1}^K\dots\sum_{k_T=1}^K\tilde\alpha_{kk_T}^T\prod_{t=2}^T\tilde\alpha_{k_tk_{t-1}}^{(t-1)}\prod_{t=1}^T F_{k_t}(y_t)$$

Since all entries in λ are strictly positive, the matrices $\tilde\Phi^{(t)}$, $t=1,\dots,T$ all have full rank. The argument in Step 2 of the proof of Theorem 3.1.1 now still applies to show that the distribution functions $H_T(\cdot;k)$, $k=1,\dots,K$, are linearly independent. □

Step 2.: If both Φ and λ have only strictly positive entries, then all parameters λ, Φ and $F_1,\dots,F_K$ are identified from the joint distribution of $\big(Y_1,\dots,Y_{2K+1}\big)$.

Proof of Step 2. It remains to identify λ. We may follow the argument in Step 3 of Theorem 3.1.1: For $T=K-1$, we may identify both $H_T(\cdot;k)$ as well as $H_{T+1}(\cdot;k)$, where we have chosen a fixed (equal) labeling for both distribution functions.

Again, we may choose $\boldsymbol{y}_1,\dots,\boldsymbol{y}_K\in\mathbb{R}^T$ such that the identified $K\times K$-matrix

$$B_1=\big(H_T(\boldsymbol{y}_t;k)\big)_{k=1,\dots,K;t=1,\dots,K}$$

has full rank K. For $y\in\mathbb{R}$ consider the identified $K\times K$-matrix

$$B_2=\big(H_{T+1}((\boldsymbol{y}_l',y)';k)\big)_{k=1,\dots,K;l=1,\dots,n}.$$

We have that

$$B_2=\tilde\Phi^{(T+1)}\operatorname{diag}\big(F_1(y),\dots,F_K(y)\big)B_1,$$

which, for y large enough so that $F_k(y) \neq 0$, $k = 1, \ldots, K$, allows to identify $\tilde{\Phi}^{(T+1)}$. Therefore, for each j, we identify

$$\frac{\tilde{\alpha}_{j,i}^{(T+1)}}{\alpha_{i,j}} = \frac{\lambda_i^{(T+1)}}{c_j}, \qquad i = 1, \ldots, K,$$

where c_j is a positive constant. If we fix j, this identifies $\lambda^{(T+1)}$ up to scale. Since $\lambda^{(T+1)}$ is a probability vector, it is itself identified and since Φ is identified and $\lambda^{(T+1)} = \lambda\Phi^T$, λ itself is identified. □

Step 3: Conclusion of the proof.
Now, we conclude the proof of the theorem. Let $t_0 = K^2 - 2K + 2$. Then from Holladay and Varga [22], Φ^{t_0} has strictly positive entries.

Observe that $\big(Y_{t_0+1}, \ldots, Y_{t_0+2K+1}\big)$ with starting vector $\lambda\Phi^{t_0}$, which has only positive entries. Using Step 1 we therefore identify Φ and $F_1, \ldots, F_K$. Then, using the result in Step 2, from

$$\big(Y_{t_0+1}, Y_{2\,t_0+1} \ldots, Y_{(2K+1)\,t_0+1}\big),$$

which is a segment of an HMM where the Markov chain starts in $\lambda\Phi^{t_0}$ and has t.p.m. Φ^{t_0}, and the state-dependent distributions are $F_1, \ldots, F_K$, we identify $\tilde{\lambda} = \lambda\Phi^{t_0}$, and therefore also $\lambda = \tilde{\lambda}\Phi^{-t_0}$.

□

Proof of Theorem 3.1.5. The case $L = K$ follows immediately from Corollary 3.1.4. Consider the case $L < K$. We add $K-L$ states which are never visited to the L-state HMM, say with state-dependent distribution equal to $\bar{F}_1$, without changing its distribution. Then from Corollary 3.1.4, we directly get a contradiction. □

Proof of Theorem 3.1.6. The existence of the limit as well as its independence from the starting distributions may be deduced from Kingman's subadditive ergodic theorem. To show definiteness, we briefly recall a construction from Leroux [30]. For a sequence (y_n) in $\mathcal{S}$, define sequences

$$u^{(n)}, v^{(n)} \in \triangle^{K-1} = \big\{(s_1, \ldots, s_K)' \in [0,1]^K : \quad s_1 + \ldots + s_K = 1\big\},$$

by

$$u_k^{(1)} = \pi_{0,k}, \quad u_k^{(n+1)} = \frac{\sum_{j=1}^K u_j^n f_{0,j}(y_n)\alpha_{0,jk}}{\sum_{j=1}^K u_j^n f_{0,j}(y_n)}, \quad k = 1, \dots, K, \;\; n = 1, 2 \dots$$
$$v_k^{(1)} = \pi_{0,k}, \quad v_k^{(n+1)} = \frac{\sum_{j=1}^K v_j^n f_j(y_n)\alpha_{jk}}{\sum_{j=1}^K v_j^n f_j(y_n)}, \quad k = 1, \dots, K, \;\; n = 1, 2 \dots .$$

where π_0 is the stationary distribution of Φ_0, and we set $0/0 = 0$. Let $\Omega = \{(y_n, u^{(n)}, v^{(n)})_{n\in\mathbb{N}}\}$. Leroux [30] shows that there is a probability measure on Ω, such that if $Q(u, v)$ denotes the distribution of $\big(u^{(1)}, v^{(1)}\big)$ under this measure, for any $T \in \mathbb{N}$ we have that

$$\begin{aligned}
&K\big((\Phi_0, f_{0,1}, \dots, f_{0,K}), (\Phi, f_1, \dots, f_K)\big)/T \\
&= \int_{(\triangle^{K-1})^2} \int_{\mathcal{S}^T} g_T(\boldsymbol{y}; u, \Phi_0, f_{0,1}, \dots, f_{0,K}) \log\Big(\frac{g_T(\boldsymbol{y}; u, \Phi_0, f_{0,1}, \dots, f_{0,K})}{g_T(\boldsymbol{y}; v, \Phi, f_1, \dots, f_K)}\Big)\, d\nu(\boldsymbol{y})\, dQ(u, v) \\
&= \int_{(\triangle^{K-1})^2} KL\big(g_T(\cdot; u, \Phi_0, f_{0,1}, \dots, f_{0,K}), g_T(\cdot; v, \Phi, f_1, \dots, f_K)\big)\, dQ(u, v).
\end{aligned}$$

Non-negativity is then obvious. To show definiteness, choose $T = (2K+1)(K^2 - 2K + 2) + 1$. Suppose that the two sets of parameters $\Phi_0, f_{0,1}, \dots, f_{0,K}$ and $\Phi, f_1, \dots, f_K$ are not equal up to label swapping. Then from Corollary 3.1.4, for any $u, v \in \triangle^{K-1}$,

$$KL\big(g_T(\cdot; u, \Phi_0, f_{0,1}, \dots, f_{0,K}), g_T(\cdot; v, \Phi, f_1, \dots, f_K)\big) > 0,$$

which immediately implies definiteness. □

4 Discussion

In the current thesis several selected aspects of the two related latent class models; finite mixtures and hidden Markov models, were considered.

The proposed combination of the EM algorithm and Newton's method for the calculation of the MLE of a multivariate Gaussian mixture performed in some special constellations better than pure EM. These constellations are characterized by a high fraction of unobserved information in the EM setting (many mixture components and a large sample size) and a low or a moderate dimension of the data. However, in the most other situations EM algorithm performed better; it failed less seldom compared to Newton's method and was faster.

Xu and Jordan [49] found a representation of the EM iteration as $\theta_{k+1} = \theta_k + P_k \nabla l(\theta_k)$, where a P_k is a well-conditioned matrix, which takes the place of the negative inverse of the Hessian $-H_k^{-1}$ in NM iterations. Hence EM can be considered as a variant of the Quasi-Newton methods. A possible subject for further research would be studying of the use of a convex combination of both matrices: $\omega_k P_k - (1 - \omega_k) H_k^{-1}$ as the iteration matrix. In doing so, one should adapt $\omega_k \in [0, 1]$ during the iterations. At the beginning ω_k should be near 1 and at the end near 0. The difficulty is to find appropriate criteria for adapting ω_k, it may depend on the condition number of the resulting matrix and/or on the negative definiteness of H_k.

Another open problem in this context is implementation and studying of Newton's method for MLE of mixtures of non-Gaussian distributions, such as t-distributions or skew-normals. In these cases, there exist no simple update formulas for all parameters for the M-step of the EM algorithm, which is why the maximization must be carried out numerically and the advantages of Newton's method should carry more weight.

A further subject of the thesis was consistency of the penalized maximum likelihood estimators for multivariate Gaussian mixtures and for univariate Gaussian hidden Markov models. The consistency proof of the penalized MLE for multivariate Gaussian mixtures from Chen and Tan [11] was elaborated and a soft spot therein

was identified and corrected with the help of a uniform law of iterated logarithm for VC-classes from Alexander [3].

A penalized maximum likelihood estimator for univariate Gaussian hidden Markov models was introduced and shown to be consistent. The proposed method consists of two stages; in the first stage a penalized mixture likelihood is maximized in order to estimate parameters of the marginal mixture and in the second stage full HMM likelihood is maximized in a neighborhood of values from the previous stage. The consistency proof generalizes the one from Chen et al. [12] for univariate Gaussian mixtures.

A possible subject for further research could be a proof of the corresponding statement for multivariate Gaussian hidden Markov models. Therefor a generalization of Alexander's uniform law of iterated logarithm for weakly dependent processes could be useful.

An alternative approach for developing penalized MLE of Gaussian HMMs could be a direct penalization of the full HMM log-likelihood, rather than that of the mixture log-likelihood as in the proposed method. But in this case, a more involved proof is required due to the analytical intractability of the HMM likelihood.

The question of identifiability of hidden Markov models with nonparametric state-dependent distributions was answered in the affirmative under the conditions that the transition matrix is ergodic and has a full rank and the state-dependent distributions are all distinct. A possible improvement of this result could be a weakening of the regularity assumption on the transition matrix. After taking a precise look at the proof, one could conjecture that the row Kruskal rank at least 2 should suffice for identification of the transition matrix and the state-dependent distributions.

Once one has identifiability of a statistical model, it is sensible to ask how to estimate this model. In the case of nonparametric hidden Markov models, the Kullback-Leibler distance of two distinct models was shown to be strictly positive by using its representation as given in Leroux [30], so the first step on the way to maximum likelihood estimation is done. In the second step one should specify a nonparametric class for the state-dependent distributions, such that each sequence of maximizers of the HMM log-likelihood based on n observations should converge a.s. to the maximizer of the negative Kullback-Leibler distance. Conditions similar to those from Wald [48] or Kiefer and Wolfowitz [27] should be imposed.

Table 4.1: **List of Notations**

e_i	The i'th unit-vector
v_k^i	The i'th element of the vector v_k
$L_k^{i,j}$	The element in the i'th row and the j'th column of the matrix L_k
$L_k^{i,\cdot}$	The i'th row of the matrix L_k
$L_k^{\cdot,i}$	The i'th column of the matrix L_k
$\vec{L}_k$	The half-vectorization of the quadratic matrix L_k (see Definition 1.2.1)
Θ	The parameter space of a statistical model
$\vec{L}_k^i$	The i'th element of the vector $\vec{L}_k$
$\vec{z}_i$	The index of the row of the i'th element of $\vec{L}_k$ in L_k
$\vec{s}_i$	The index of the column of the i'th element of $\vec{L}_k$ in L_k
$\|L\|$	The absolute value of the determinant of the matrix L_k
$'$	The transpose operator
$\nabla_\theta l$	The gradient of the function l w.r.t. θ
$\nabla_\theta^2 l$	The Hessian of the function l w.r.t. θ
$\|\|v\|\|$	The euclidean norm of the vector v
$\delta_i(j)$	Kronecker delta
$\mathrm{diag}(v)$	For a vector v: a diagonal matrix with elements of v on the diagonal
$\mathrm{diag}(L)$	For a matrix L: the diagonal elements of M as a vector
$\mathbb{N}$	The set of natural numbers
$\mathbb{Z}$	The set of integers
$\mathbb{R}$	The set of real numbers
$\mathbb{R}^{d\times d}$	The set of real $d \times d$ matrices
$\mathbb{R}^d$	The set real d-vectors
$\mathbb{R}_{lt}^{d\times d}$	The set of real lower triangular $d \times d$ matrices
$\mathcal{P}^d$	The set of $d \times d$ symmetric positive matrices
S^{d-1}	$d-1$-sphere
$\mathbb{E}_0$	Expectation w.r.t. true parameter θ_0
$d_c(x,y)$	$\sum_{s=1}^r \|\arctan(x_s) - \arctan(y_s)\|$ metric on $\mathbb{R}^r$
$\mathcal{T}$	The set of transition probability matrices
$\triangle^{K-1}$	$\{(\alpha_1,\ldots,\alpha_K) \in \mathbb{R}^K,\ \sum_{i=1}^K \alpha_i = 1,\ \alpha_i \geq 0\}$
$\xrightarrow{d}$	Convergence in distribution
$\xrightarrow{P}$	Convergence in probability

Bibliography

[1] Aitkin, M. and Aitkin, I. (1996). A hybrid EM/Gauss-Newton algorithm for maximum likelihood in mixture distributions. *Statistics and Computing*, **6** 127–130.

[2] Akama, Y. and Irie, K. (2011). VC dimension of ellipsoids. *ArXiv e-prints*. 1109.4347.

[3] Alexander, K. S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Annals of Probability*, **12** 1041–1067.

[4] Alexandrovich, G. (2014). A note on the article 'inference for multivariate normal mixtures' by j. chen and x. tan. *Journal of Multivariate Analysis*, **129** 245 – 248. URL http://www.sciencedirect.com/science/article/pii/S0047259X14000827.

[5] Alexandrovich, G. and Holzmann, H. (2014). Nonparametric identification of hidden Markov models. *ArXiv e-prints*. 1404.4210.

[6] Allman, E. S., Matias, C. and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, **37** 3099–3132.

[7] Baudry, J. P., Raftery, A. E., Celeux, G., Kenneth, L. and Gottardo, R. (2008). Combining mixture components for clustering. *Inria, Rapport de recherche no. 6644*.

[8] Bickel, J. P., Ritov, Y. and Rydèn, T. (1998). Asymptotic Normality of the maximum likelihood estimator for general Hidden Markov Models. *The Annals of Statistics*, **26** 1614–1635.

[9] Billingsley, P. (1986). *Probability and Measure*. John Wiley & Sons.

[10] Celeux, G., Chauveau, D. and Diebolt, J. (1995). On stochastic versions of the em algorithm. *Inria, Rapport de recherche no. 2514*.

[11] Chen, J. and Tan, X. (2009). Inference for multivariate normal mixtures. *Journal of Multivariate Analysis*, **100** 1367–1383.

[12] Chen, J., Tan, X. and Zhang, R. (2008). Inference for normal mixtures in mean and variance. *Statistica Sinica*, **18** 443–465.

[13] Ciuperca, G., Ridolfi, A. and Idier, J. (2003). Penalized maximum likelihood estimator for normal mixtures. *Scandinavian Journal of Statistics*, **30** 645–59.

[14] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, **39** 1–38.

[15] Everitt, B. S. (1984). Maximum Likelihood Estimation of the Parameters in a Mixture of Two Univariate Normal Distributions; A Comparison of Different Algorithms. *Journal of the Royal Statistical Society. Series D*, **33** 205–215.

[16] Ferguson, T. S. (1996). *A course in large sample theory*. Chapman & Hall.

[17] Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97** 611–631.

[18] Fraley, C. and Raftery, A. E. (2006). Mclust version 3 for r normal mixture modeling and model-based clustering.

[19] Gassiat, E., Cleynen, A. and Robin, S. (2013). Finite state space non parametric hidden markov models are in general identifiable. *preprint*.

[20] Hathaway, R. J. (1985). A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions. *Annals of Statistics*, **13** 795–800.

[21] Hennig, C. (2010). Methods for merging gaussian mixture components. *Advances in Data Analysis and Classification*. URL `http://dx.doi.org/10.1007/s11634-010-0058-3`.

[22] Holladay, J. C. and Varga, R. S. (1958). On powers of non-negative matrices. *Proceedings of American Mathematical Society*.

[23] Jakowitz, S. J. and Spargins, J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics* 209–214.

[24] Jamshidian, M. and Jennrich, I. R. (1997). Acceleration of the EM Algorithm by Using Quasi-Newton Methods. *Journal of the Royal Statistical Society. Series B*, **59** 569–587.

[25] Jank, W. (2006). The EM Algorithm, Its Stochastic Implementation and Global Optimization Some Challenges and Opportunities for OR. *Inria, Rapport de recherche no. 2514*.

[26] Kelley, C. T. (1995). Iterative methods for linear and nonlinear equations. *SIAM Publications*.

[27] Kiefer, J. and Wolfowitz, J. (1956). Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. *Annals of Mathematical Statistics*, **27** 887–906.

[28] Kruskal, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications* 95–138.

[29] Lange, K. (1995). A quasi-Newton acceleration of the EM algorithm. *Statistica Sinica*, **5** 1–18.

[30] Leroux, B. G. (1990). Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications*, **40** 127–143.

[31] Lindgren, G. (1978). Markov Regime Models for Mixed Distributions and Switching Regressions. *Scand. J. Statistics* 81–91.

[32] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

[33] McLachlan, G. J. and Bashford, K. E. (1998). *Mixture Models Inference and Applications to Clustering*. Marcel Dekker, New York.

[34] McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Willey.

[35] Merlevède, F., Peligrad, M. and Rio, E. (2011). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, **151** 435–474.

[36] Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. S Springer.

[37] Norris, J. (1998). *Markov Chains*. No. Nr. 2008 in Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press. URL `http://books.google.de/books?id=qM65VRmOJZAC`.

[38] Peters, B. C. and Walker, H. F. (1978). An Iterative Procedure for Obtaining Maximum-Likelihood Estimates of the Parameters for a Mixture of Normal Distributions. *SIAM Journal on Applied Mathematics*, **25** 362–378.

[39] Petrie, T. (1969). Probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, **40** 97–115.

[40] R. Lebret, F. L., S. Ioveff (2013). Rmixmod an interface of MIXMOD, v.1.1.3.

[41] Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of Speech Recognistion*. Prentice-Hall International, Inc.

[42] Redner, R. A. and Walker, H. F. (1984). Mixture Densities, Maximum Likelihood and the Em Algorithm. *SIAM Review*, **26** 195–239.

[43] Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Willey, New York.

[44] Tanaka, K. (2009). Strong consistency of the maximum likelihood estimator for finite mixtures of location-scale distributions when penalty is imposed on the ratios of the scale parameters. *Scandinavian Journal of Statistics*, **36** 171–184.

[45] Teicher, H. (1967). Identifiability of mixtures of product measures. *Ann. Math. Stat.*, **38** 1300–1302. URL `http://dx.doi.org/10.1214/aoms/1177698805`.

[46] Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons.

[47] van der Vaart, A. and Wellner, J. A. (2000). *Weak Convergence and Empirical Processes*. Springer.

[48] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, **20** 595–601.

[49] Xu, L. and Jordan, M. I. (1995). On convergence properties of the em algorithm for gaussian mixtures. *Neural Computation*, **8** 129–151.

[50] Zucchini, W. and MacDonald, I. (2009). *Hidden Markov Models for Time Series*. Chapman & Hall.